Università degli Studi di Padova

Dipartimento di Scienze Statistiche

Corso di Laurea Magistrale in

Scienze Statistiche

Analisi dell'impatto di fattori ambientali sugli accessi in pronto soccorso per cause legate alle alte vie respiratorie

Relatore Prof. Bruno Scarpa

Laureanda Elisa Masetto

Matricola N. 2039334

Anno Accademico 2022/23

Indice

Capitolo 1

Introduzione

1.1 Qualità dell'aria e salute

Oltre il 70% della popolazione europea vive in ambienti urbanizzati, in cui il verde urbano è uno dei principali strumenti per intervenire sull'equilibrio dell'ecosistema cittadino. Tuttavia, l'impiego di piante arboree e arbustive in grado di incrementare la concentrazione di pollini e spore in città è tra le cause dell'aumento negli ultimi anni di manifestazioni allergiche, che colpiscono maggiormente i soggetti che vivono in areee urbane rispetto alle zone rurali [Ortolani et al., 2015]. Per spiegare il progressivo aumento delle malattie allergiche si è ipotizzato che la responsabilità possa essere attribuita anche allo stile di vita urbano, caratterizzato dall'uso frequente di veicoli ad emissioni di scarico e soggetto generalmente all'esposizione ad inquinamento atmosferico. Molti studi epidemiologici hanno dimostrato le ripercussioni negative sulla salute umana provocate dall'inquinamento atmosferico [D'Amato et al., 2010]. In aggiunta, quest'ultimo può incrementare l'attività biologica allergenica del polline, come è stato dimostrato da recenti ricerche. Il polline, infatti, sarebbe in grado di catturare e trasportare nelle vie respiratorie i principali inquinanti atmosferici, quali ozono (O_3), il biossido di azoto (NO_2) e il particolato (polveri sottili), causando nei soggetti sensibilizzati manifestazioni di ipersensibilità agli allergeni, e innescando reazioni allergiche anche nei soggetti non allergici [Reinmuth-Selzle et al., 2023]. Inoltre, l'aumento della temperatura terreste e il cam-

biamento climatico interferiscono con la produzione stagionale di pollini allergenici, prolungandone i periodi di presenza nell'aria, oltre che a interagire con la presenza e l'accumulo degli inquinanti.

L'inquinamento atmosferico continua ad avere impatti significativi sulla salute della popolazione europea, in particolare nelle aree urbane, e colpisce maggiormente alcuni gruppi di popolazioni più suscettibili, come gli anziani, i bambini e le persone con condizioni di salute già compromesse. Il rapporto del 2020 sulla qualità dell'aria in Europa dell'Agenzia Europea dell'Ambiente (EEA, *European Enviroment Agency*) fornisce alcune stime dell'impatto sulla salute dell'esposizione all'inquinamento atmosferico. In particolare, viene indicano che nel 2018 l'esposizione a lungo termine al particolato PM2.5 in Europa è stato responsabile di circa 417.000 morti premature. L'impatto stimato attribuibile all'esposizione della popolazione al biossido di azoto è stato invece di circa 55.000 morti premature, mentre si stima che l'esposizione all'ozono a livello del suolo abbia causato 20.600 morti premature nel 2018 in Europa. L'inquinamento atmosferico danneggia anche la vegetazione e gli ecosistemi, nonché la qualità dell'acqua e del suolo [European Environmental Agency, 2020; SNPAmbiente, 2020].

Nonostante le misure sempre più severe adottate dall'Unione Europea per far fronte all'inquinamento atmosferico, gli standard attuali sulla qualità dell'aria sono ancora violati in alcune località europee, tra cui la Pianura Padana. Oltre infatti al complesso scenario emissivo dovuto all'alta densità di industrie nel nord Italia, le caratteristiche topografiche della Pianura Padana, come la presenza di catene montuose e la carenza di piogge, favoriscono il ristagno atmosferico e l'accumulo di sostanze inquinanti. Il *lockdown* del 2020 dovuto alla pandemia da Covid-19 rappresenta uno scenario unico in cui alcuni tipi di emissioni, specialmente quelle da traffico urbano, sono in gran parte cessate. In Italia, rispetto al periodo pre-Covid, il traffico di veicoli leggeri ha subito una riduzione dell'80%, e del 40% per i veicoli pesanti. Sembra comunque che gli effetti del *lockdown* del 2020 siano stati principalmente legati alla riduzione degli ossidi di azoto. Anche le concentrazioni di PM10 sono state complessivamente inferiori in tutta Europa, sebbene l'impatto fosse meno pronunciato rispetto al biossido di azoto, suggerendo che il traffico veicolare non sia la principale fonte di emissione delle polveri sottili che sono invece causate maggiormente dai processi di combustione delle attività industriali [Pivato et al., 2023].

Il Veneto, collocato nella parte orientale della Pianura Padana, risulta una delle regioni più inquinate in Italia. Nel 2021 gran parte delle città di provincia, tra cui la città di Padova, figuravano tra quelle con le più alte concentrazioni medie annuali di particolato, mentre nel 2022 Padova appare come una delle città maggiormente inquinate da biossido di azoto [Legambiente, 2022, 2023].

L'obiettivo del presente eleborato è studiare l'impatto di fattori ambientali, quali allergeni, inquinamento atmosferico e variabili climatiche, sulla salute in termini di accessi registrati al pronto soccorso dell'ospedale di Padova dovuti a patologie legate alle alte vie respiratorie. Il software utilizzato per l'analisi statistica è `R`.

Nelle sezioni seguenti di questo capitolo vengono descritti i fattori ambientali considerati nello studio e le patologie di interesse analizzate.

1.2 Inquinamento atmosferico

Gli inquinanti considerati sono le polveri sottili, l'ozono e il biossido di azoto.

Polveri sottili. Le polveri sottili (particolato, *Particular Matter*, PM) sono una miscela di solidi organici e inorganici e particelle liquide di diversa origine, dimensione e composizione. Sono caratterizzate da lunghi tempi di permanenza in atmosfera, e possono essere trasportate anche a lunghe distanze dal punto di emissione. Il particolato viene distinto in tre classi, a seconda della dimensione del diametro aerodinamico: PM10, PM2.5, PM1. Le particelle con diametro inferiore a 10 µm (PM10, anche detto particolato grossolano) possono penetrare nelle basse vie aeree, mentre il particolato fine, cioè le particelle con diametro inferiore o uguale a 2,5 µm (PM2.5), può essere inalato più profondamente nei polmoni. Il particolato è stato significativamente associato alle visite al pronto soccorso a causa di sintomi del tratto respiratorio inferiore, nonché all'uso di farmaci anti-asma e visite mediche per l'asma [D'Amato et al., 2010]. Le polveri sottili possono avere origine da fenomeni naturali (processi di erosione del suolo, incendi boschivi, ecc.), ma soprattutto da attività antropiche, in particolar modo dai processi di combustione e dal traffico veicolare.

Per le polveri fini PM2.5, in Italia è stabilito che il valore medio annuale per la protezione della salute umana non dovrebbe superare i 25 $\mu g/m^3$. Per le polveri PM10 viene stabilito anche un

valore medio giornaliero, 50 µg/m^3, che non dovrebbe essere superato più di 35 volte l'anno. La soglia media annuale è 40 µg/m^3. L'Organizzazione Mondiale della Sanità raccomanda di abbassare il valore soglia giornaliero e annuale rispettivamente a 45 µg/m^3 e 15 µg/m^3 per le PM10, e a 5 µg/m^3 per le PM2.5 [Organizzazione Mondiale della Sanità, 2022].

In Veneto nel 2021 erano presenti 39 stazioni di rilevazioni del particolato. Di queste, 29 stazioni (il 74%) hanno registrato il superamento del valore soglia giornaliero di 50 µg/m^3 più di 35 volte l'anno, evidenziando maggiori criticità per le zone di pianura [ARPAV, 2021].

L'Agenzia Regionale per la Prevenzione e Protezione Ambientale del Veneto (ARPAV) segnala livelli di allerta in base alla concentrazione misurata in giorni successivi. Il primo livello di allerta (arancione) si attiva dopo 4 giorni consecutivi in cui viene superato il limite giornaliero di 50 µg/m^3. Se gli sforamenti si protraggono per oltre 10 giorni consecutivi si attiva il secondo livello di allerta (rosso) [ARPAV, 2023].

Ozono. L'ozono è un gas formato da tre atomi di ossigeno (O_3) e costituisce, nella stratosfera, uno strato protettivo dalle radiazioni ultraviolette provenienti dal sole, rendendosi dunque indispensabile per la vita sulla terra. Negli strati bassi dell'atmosfera (troposfera) esso è presente in basse concentrazioni. Tuttavia, la presenza di alcuni inquinanti chimici, in concomitanza di fattori meteo-climatici favorevoli (alte temperature estive), può aumentarne la concentrazione, causando l'inquinamento da ozono. Gli inquinanti "precursori" sono generalmente di tipo antropico, e il loro accumulo e ristagno può essere influenzato da diverse variabili meteorologiche come l'intensità della radiazione solare, la temperatura, la direzione e la velocità del vento. La presenza di elevati livelli di ozono può causare irritazioni agli occhi, naso, gola e apparato respiratorio, oltre ad essere dannosa anche per la salute degli animali e delle piante. L'ozono a livello del suolo può infatti danneggiare colture, foreste e altra vegetazione, compromettendone la crescita e portando cambiamenti nella biodiversità.

In Italia, il livello orario da mantenere al fine di evitare, prevenire o ridurre effetti nocivi per la salute umana e l'ambiente è di 120 µg/m^3, mentre il superamento di 180 µg/m^3 sussiste un rischio per la salute umana in caso di esposizione di breve durata per alcuni gruppi particolarmente sensibili della popolazione e impone di adottare provvedimenti tempestivi [ARPAV, 2023].

Biossido di azoto. Il biossido di azoto (NO_2) è un inquinante che si forma in seguito a processi di combustione, generati principalmente dal traffico veicolare (in particolare dai motori diesel), negli impianti industriali, negli impianti di produzione di energia elettrica, di riscaldamento civile e di incenerimento dei rifiuti [SNPAmbiente, 2020]. Numerose ricerche hanno evidenziato un'associazione statisticamente significativa tra le concentrazioni atmosferiche giornaliere di NO_2 e le consultazioni mediche, i ricoveri ospedalieri per malattie respiratorie e l'incidenza di attacchi d'asma [Latza et al., 2009]. In Italia, il valore limite medio annuale di biossido di azoto per la protezione della salute umana è di 40 $\mu g/m^3$ [ARPAV, 2022].

1.3 Allergeni

Gli allergeni considerati in questo studio sono le muffe prodotte dall'alternaria e i pollini generati dalla fioritura delle piante appartenenti alle famiglie delle betulacee, composite, coriacee e graminacee. Segue una breve descrizione delle specie.

Alternaria. L'alternaria è un genere di fungo la cui muffa prolifera in ambienti con un tasso di umidità generalmente superiore al 65% e ad una temperatura compresa tra i 18 e 32 gradi. In un paese dal clima temperato come l'Italia, le spore di alternaria possono rimanere nell'aria nei mesi che vanno da maggio a novembre, con picchi tra la fine dell'estate e l'inizio dell'autunno. L'alternaria compare tipicamente sulle foglie di molte piante, sulla frutta e sul legno umido. Le sue spore sono facilmente inalabili a causa della loro ridotta dimensione, causando sintomi quali rinite, congiuntivite e asma.

Betulacee. Le betulacee, presenti soprattutto nel nord Europa ma diffuse anche in Italia, comprendono le piante di betulla e ontano bianco [Blackmore et al., 2003]. Per entrambe, il periodo di dispersione del polline è tra febbraio e aprile. La betulla è di frequente presente nelle città e nei parchi pubblici ed è responsabile di sintomatologia oculo-rinitica e asma, così come l'ontano [Ortolani et al., 2015].

Composite. La famiglia delle Composite (*asteraceae* contiene il maggior numero di specie riconosciute e descritte (circa 24 mila), che crescono in quasi tutti i tipi di ambiente [Funk et al.,

2009]. Le Composite che causano sensibilizzazioni di tipo allergico sono relativamente poche, in rapporto al grande numero di componenti di questa famiglia vegetale, e la più rilevante è l'ambrosia. L'Ambrosia è un'erba infestante importata accidentalmente in Europa dagli Stati Uniti alla fine del diciannovesimo secolo. Negli ultimi decenni la prevalenza di ambrosia è aumentata in Europa, specialmente in alcune zone tra cui il nord Italia, in cui la percentuale di pazienti sensibilizzati a questo allergene è passata dal 24% al 70% nel ventennio 1989-2008. La fioritura è tra agosto e settembre. I pollini rilasciati causano tipicamente sintomi quali rinite, congiuntivite e asma [de Weger et al., 2016].

Coriacee. Le coriacee comprendono gli alberi di carpino bianco, carpino nero e nocciolo. Il potenziale allergenico è analogo a quello delle betulacee, così come il periodo di dispersione del polline (ad eccezione del carpino nero, il cui periodo è aprile-maggio). I sintomi dell'allergia al nocciolo tendono spesso a perdurare nel tempo a causa della reazione crociata con i pollini di ontano e betulla; i pazienti allergici a questi pollini possono inoltre presentare sindrome allergica orale in seguito all'ingestione di alcuni alimenti, tra cui mele e noci [Ortolani et al., 2015].

Graminacee. Le graminacee (*poaceae*) sono piante erbacee che includono circa 12 mila specie. Diffuse in ambienti erbacei come pascoli, prati, terreni coltivati ed incolti, coprono buona parte della superficie terrestre e rivestono una notevole importanza ecologica. Producono grandi quantità di polline, che viene rilasciato nell'atmosfera ed è attualmente classificato tra i principali aeroallergeni. Il polline è presente nell'aria nei mesi primaverili-estivi, ed è responsabile di sintomi allergici, quali rinocongiuntivite e asma, soprattutto nel nord America e in Europa. Nei soggetti sensibilizzati può inoltre predisporre alla sindrome allergica orale [García-Mozo, 2017].

La *Rete Italiana di Monitoraggio Aerobiologico* fornisce la suddivisione della concentrazione di spore fungine e pollini in quattro classi (assente, bassa, media, alta), sottolineando però che tali valori non costituiscono soglie di rischio allergia (Tab.1.1). Nonostante le varie ricerche scientifiche, infatti, non si è ancora riusciti ad individuare con certezza i valori soglia per le concentrazioni al di sopra dei quali si manifestano i sintomi allergici. Queste infatti possono dipendere dalle aeree geografiche, dalle condizioni meteorologiche e dall'inquinamento. Inoltre, la concentrazione soglia può variare non solo da soggetto a soggetto, ma può cambiare per uno stesso soggetto durante la

Concentrazione	Assente	Bassa	Media	Alta
Spore fungine				
Alternaria	0-1	1-10	10-100	>100
Pollini				
Betulacee	0-0.5	0.5-16	16-50	>50
Composite	0-0.1	0.1-5	5-25	>25
Coriacee	0-0.5	0.5-16	16-50	>50
Graminacee	0-0.5	0.5-10	10-30	>30

Table 1.1: Livelli di concentrazione di spore e pollini (g/m^3).

stagione di pollinazione [Rete Italiana di Monitoraggio Aerobiologico, 2022]. Tuttavia, numerosi studi in letteratura sostengono che una concentrazione di almeno 100 spore/m^3 di alternaria è in grado di causare un'ampia gamma di reazioni allergiche negli individui sensibilizzati [Jones, 2023]. La stessa evidenza non è riscontrabile per le concentrazioni di polline, ma si possono comunque individuare alcune linee guida. Revisioni sistematiche presenti in letteratura hanno individuato per le betulacee una concentrazione soglia di 45 grammi/m^3 per la comparsa di primi sintomi, e di 75-85 grammi/m^3 per la manifestazione di reazioni allergiche in tutti i soggetti sensibilizzati [Steckling-Muschack et al., 2021; Ojrzyńska et al., 2020]. L'ambrosia causa sintomi allergici nei soggetti sensibilizzati già a basse concentrazioni: si riscontra infatti un valore soglia tra i 10 e i 20 grammi/m^3 [Tosi et al., 2011; Buters et al., 2015]. Il valore soglia per le coriacee risulta 35 grammi/m^3 [Nowosad et al., 2016], mentre per le graminacee è 50 grammi/m^3 [Piotrowska-Weryszko and Weryszko-Chmielewska, 2014].

1.4 Clima

I fattori climatici considerati nello studio sono la temperatura, le precipitazioni, l'umidità minima e massima giornaliera.

Temperatura. Viene considerato il valore medio giornaliero, misurato in gradi Celsius.

Precipitazioni. La quantità di pioggia caduta viene misurata in millimetri dai pluviometri,

strumenti installati a un'altezza dal suolo di circa un metro e mezzo in luoghi aperti, lontani da alberi e da fabbricati, in modo che la pioggia sia libera di cadere sul ricevitore. 1 mm di precipitazione equivale ad 1 litro su un metro quadro di superficie. Pertanto, dire ad esempio che la quantità di pioggia caduta in una certa località è di 20 mm, equivale a dire che su ogni area di 1 metro quadrato in quella località sono caduti 20 litri di pioggia. I millimetri di pioggia caduti in un'ora ne definiscono l'intensità: un valore maggiore di 6 mm/h è indicativo di pioggia forte o molto forte (rovescio, nubifragio) [CNR, 2021].

Umidità relativa. L'umidità relativa è il rapporto, espresso in percentuale, tra la quantità di vapore d'acqua presente nella massa d'aria e la quantità massima che essa può contenere a quella temperatura ed alla stessa pressione. L'umidità relativa si misura con l'igrometro, posizionato ad un'altezza di circa 2 metri dal suolo.

1.5 Patologie legate alle alte vie respiratorie

Le cause di accesso al pronto soccorso esaminate in relazione ai fattori ambientali sono patologie che fanno riferimento alla branca della medicina otorinolaringoiatrica, quali laringite, faringite, tracheite, tonsillite, otite media e rinosinusite. Queste patologie sono legate ad una qualche compromissione delle alte vie respiratorie, ossia la parte dell'apparato respiratorio che costituisce la prima via di immissione dell'aria ambientale.

Laringite. Si tratta di un'infiammazione della laringe che si manifesta con quello che comunemente viene chiamato "mal di gola", e comporta la sensazione di bruciore e dolore alla parte posteriore del cavo orale, oltre a un senso di fastidio e difficoltà nella deglutizione, e in alcuni casi febbre. La laringite è correlata a un'infezione del tratto respiratorio superiore ed è provocata da una varietà di cause, tra cui l'esposizione ad aria fredda o smog, così come da infezioni virali come raffreddore o influenza.

Faringite. La faringite è un'infiammazione della faringe, causata principalmente da infezioni virali, e presenta sintomi comuni alla laringite. Spesso faringe e laringe vengono colpite insieme, e si parla pertanto di *laringofaringite*.

Tracheite. La tracheite è un'infiammazione a carico della trachea, provocata principalmente da un'infezione di causa batterica, ed è tipicamente associata a tosse, difficoltà respiratorie e febbre.

Tonsillite. Con tonsillite si intende l'infiammazione, generalmente di natura virale, delle tonsille, con conseguente ingrandimento delle stesse e con dolore riferito alla gola e in qualche caso all'orecchio.

Otite media. L'otite media è un'infezione batterica o virale dell'orecchio medio che in genere accompagna un'infezione delle vie aeree superiori. È infatti una comune complicanza di raffreddore, faringite, influenza ed allergie.

Rinosinusite. La rinosinusite è un processo infiammatorio che coinvolge il naso e le cavità paranasali. Il processo infiammatorio causa un accumulo di muco che rende difficoltosa la respirazione e crea un terreno ottimale per lo sviluppo batterico. La rinosinuite può essere di origine virale (il comune raffreddore), o di origine batterica, e può comportare vari sintomi tra cui congestione nasale, tosse, riduzione del gusto e dell'olfatto, febbre.

1.6 Impatto del Covid-19 sugli accessi in pronto soccorso

Le misure restrittive adottate per affrontare la pandemia da Covid-19 hanno comportato, oltre che ad un caso eccezionale di riduzione di alcuni tipi di inquinanti, anche una rilevante riduzione degli accessi in pronto soccorso. Come dimostrato da studi retrospettivi condotti nel nord Italia, nel periodo pandemico gli ospedali hanno registrato un picco di accessi per ricoveri Covid-correlati, ai quali si è però accompagnato un drastico decremento negli accessi e ricoveri in pronto soccorso per altre patologie. La paura di infettarsi avrebbe colpito in particolar modo soggetti che avrebbero probabilmente richiesto cure di minore intensità (codici bianchi e verdi). È stata osservata anche una riduzione dei traumi, riconducibile al fatto che il *lockdown* abbia nettamente ridotto il rischio di incorrere in eventi traumatici, quali incidenti stradali, sportivi o lavorativi [Giostra et al., 2021].

Capitolo 2

Dati

2.1 Accessi in Pronto Soccorso

I dati analizzati in questo studio sono stati forniti dal Pronto Soccorso dell'Azienda Ospedaliera di Padova in riferimento a due periodi di tempo: dal 1° gennaio al 31 dicembre 2017, e dal 9 marzo 2020 all'8 marzo 2021. Entrambi i periodi coprono dunque la durata di un anno e fanno riferimento a due circostanze diverse. Il secondo periodo corrisponde infatti al primo anno di pandemia causata dall'infezione da Covid-19, in cui le restrizioni e il *lockdown* limitavano, soprattutto nei primi mesi, gli spostamenti, le attività e il traffico cittadino (il 9 marzo 2020 nella città di Padova, così come in tutta Italia, cominciava il primo *lockdown*). Il primo periodo, quello del 2017, si riferisce invece a una situazione di normalità. Per entrambi i periodi, i dati forniti constano di un dataset in cui l'unità statistica è il singolo accesso al pronto soccorso. Per ogni paziente registrato vengono rilevate le variabili di sesso, età e diagnosi principale. In riferimento ai dati del 2020, la diagnosi si può ritenere svincolata dall'infezione da Covid-19, perché sono stati considerati solo i pazienti che sono risultati negativi al tampone effettuato all'ingresso al pronto soccorso. Viene riportata inoltre la data di accesso in pronto soccorso. Ai fini dell'analisi, sono stati mantenuti solo i pazienti la cui diagnosi principale era una delle patologie legate alle alte vie respiratorie descritte nel capitolo precedente (che verranno chiamati "casi"). Sono stati mantenuti inoltre tutti i pazienti la cui causa

di accesso era un qualche tipo di trauma (facciale, auricolare, laringeo, nasale, presenza di corpi estranei), in modo da avere un gruppo di controllo. Le numerosità totali ottenute per il 2017 e il 2020 risultavano rispettivamente 2368 e 1457.

2.1.1 Analisi descrittive

In tabella 2.1 e 2.2 viene mostrata la statistica descrittiva degli accessi in pronto soccorso. Si osserva che, a parità di lunghezza dei periodi considerati, nel 2020 si registra un numero inferiore di accessi totali per patologia rispetto al 2017. Si passa infatti da un totale di 1502 accessi nel 2017 a 726 nel periodo 2020-2021, con una diminuzione del 52%. Diminuisce anche la percentuale di accessi per patologia sul totale degli accessi considerati: nel 2017, sul totale dei 2368 accessi, il 62% era per causa specifica, mentre nel 2020 solo il 50% degli accessi considerati è dovuto a una patologia. Considerando solo gli accessi per patologia (tab. 2.2), si osserva che nel 2017 la patologia che causa più accessi in pronto soccorso è l'otite media (29% di accessi), la cui percentuale si riduce al 22% nel 2020 ed è preceduta solo dalla faringite (37%). In generale, nel periodo di pandemia si osserva una diminuzione degli accessi in pronto soccorso che può essere determinata anche dalle misure restrittive, oltre che dal timore da parte dei pazienti di accedere ad aree maggiormente esposte al rischio di infezione come appunto il pronto soccorso e l'ospedale. Inoltre, la pandemia ha portato le persone a trascorrere più tempo al chiuso, indossare mascherine all'aperto e a lavarsi le mani più frequentemente: questi comportamenti hanno contribuito a limitare l'esposizione e il contatto con le sostanze presenti nell'aria. In figura 2.1 viene riportata la differenza percentuale sul numero di accessi totali per patologia (casi) e sul numero di accessi per trauma (controlli), distinta per mesi. Gli accessi per patologie non Covid-correlate subiscono una netta riduzione specialmente nei mesi di maggiore crisi pandemica. La riduzione dei traumi è consistente soprattutto nel primo periodo di *lockdown*, mentre è praticamente nulla nei periodi di maggiore allentamento delle restrizioni.

	2017	**2020**
	N = 2368	N = 1457
Età	36 (21, 57)	41 (21, 64)
Sesso		
Femmina	1156 (49%)	690 (47%)
Maschio	1212 (51%)	767 (53%)
Diagnosi		
Faringite	210 (8.7%)	272 (19%)
Laringite	273 (11%)	91 (6.2%)
Otite media	433 (18%)	158 (11%)
Rinosinusite	229 (9.5%)	84 (5.8%)
Tonsillite	353 (15%)	112 (7.7%)
Tracheite	4 (0.2%)	9 (0.6%)
Trauma	910 (38%)	731 (50%)

Table 2.1: Mediana (IQR) per età; distribuzione assoluta (percentuale) per sesso e diagnosi principale di accesso in pronto soccorso.

Patologia	2017	2020
Faringite	210 (14%)	272 (37%)
Laringite	273 (18%)	91 (13%)
Otite media	433 (29%)	158 (22%)
Rinosinusite	229 (15%)	84 (16%)
Tonsillite	353 (24%)	112 (15%)
Tracheite	4 (0.3%)	9 (1.2%)
Totale	1502 (100%)	726 (100%)

Table 2.2: Distribuzione assoluta (percentuale) per la patologia di accesso in pronto soccorso.

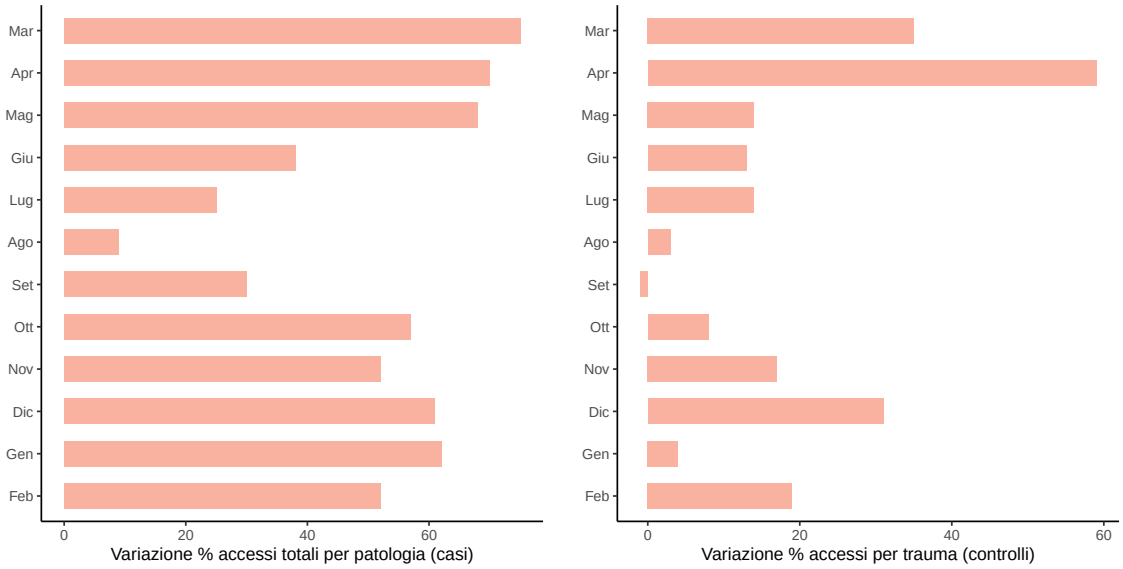


Figure 2.1: Variazione percentuale per mese tra gli accessi del 2017 e quelli del 2020.

2.2 Variabili ambientali

Accanto alle variabili di singolo accesso, sono state affiancate le variabili giornaliere riferite ai fattori ambientali. Dall'ARPAV sono dunque state reperite le informazioni sulla concentrazione media giornaliera degli allergeni (in spore o grammi per metro cubo), la temperatura media (in °C), l'umidità relativa massima e minima (in %), il livello medio di precipitazioni (in mm), la concentrazione degli inquinanti (in milligrammi su metro cubo). Vediamo di seguito alcune operazioni preliminari sui dati.

Dati 2020. Per tutti gli allergeni, la concentrazione media giornaliera non è stata registrata dal 20 al 26 luglio e dal 7 dicembre 2020 al 10 gennaio 2021, ad eccezione per le coriacee il cui dato risulta mancante solo dal 7 dicembre al 10 gennaio. Si è deciso di imputare questi dati osservando gli andamenti temporali delle concentrazioni degli allergeni, riportati in figura 2.2. In particolare, si osserva che per betulacee, composite, coriacee i dati mancanti si possono imputare a 0, poiché la concentrazione giornaliera di questi allergeni risultava nulla nelle date adiacenti a quelle mancanti. Un discorso analogo si può fare per i dati mancanti dell'alternaria e delle graminacee nel periodo

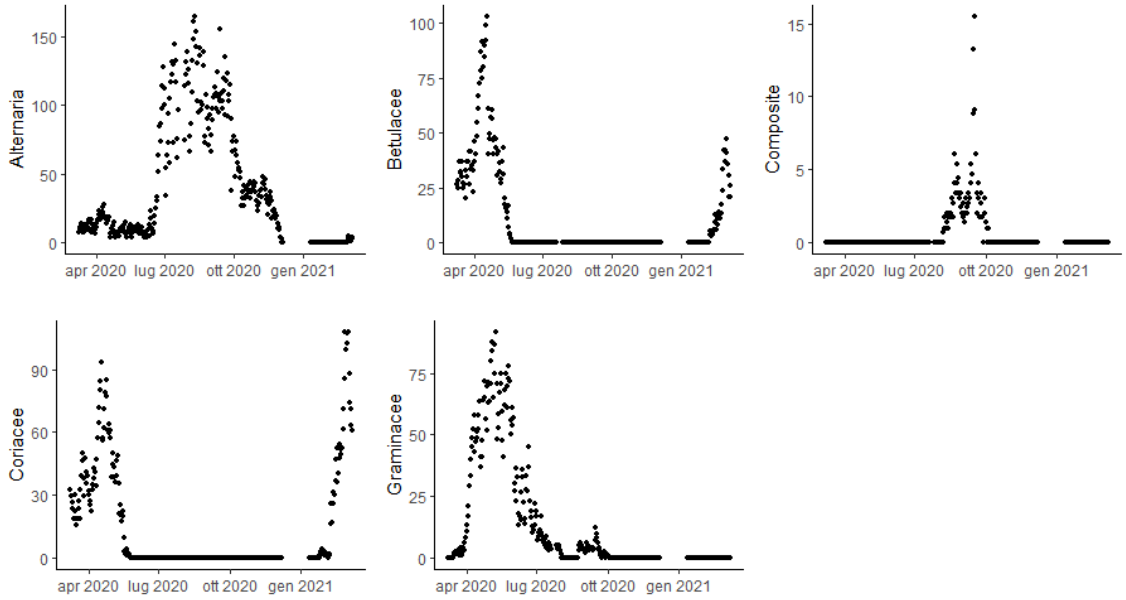
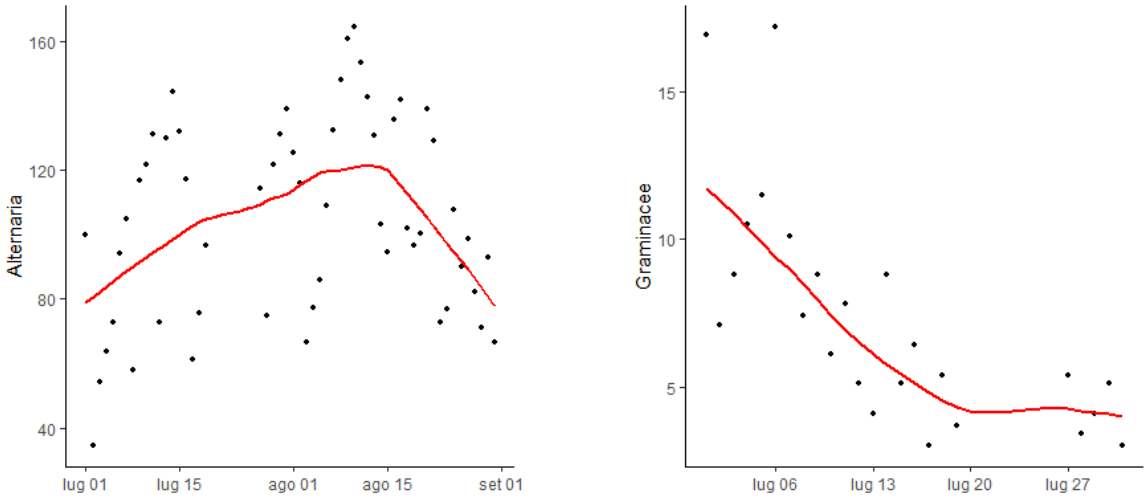


Figure 2.2: Andamento temporale della concentrazione degli allergeni nel 2020.

invernale, mentre in quello estivo è stata effettuata un'imputazione tramite regressione lineare locale con il metodo *loess* (*Locally Estimated Scatter plot Smoothing*), impostando il parametro di *span* (frazione di punti considerati nell'intorno) pari a 0.6 (fig. 2.3).

Per misurare il particolato (PM10), sono stati reperiti i valori medi giornalieri di due stazioni di rilevamento, una situata a sud rispetto al centro della città di Padova (Mandria) e una al nord (Arcella). Anche in questo caso, in alcune giornate le concentrazioni non sono state rilevate. Per la stazione di Mandria si ha il 2.5% di dati mancanti (corrispondente a 9 giornate), mentre per Arcella il 2.7% (10 giornate). La correlazione tra le due rilevazioni, calcolata sulle osservazioni complete, è pari a 0.97, ed indica una forte associazione tra le due misurazioni. Questo rende possibile usare solo una delle due, imputando i dati mancanti tramite una regressione lineare tra le due misurazioni. In questo caso, sono stati utilizzati i dati della stazione di Arcella per imputare quelli di Mandria. La relazione lineare stimata che lega il particolato di Mandria a quello di Arcella (fig. 2.4) risulta

$$PM10_{Mandria} = 0.864 + 0.845 \cdot PM10_{Arcella} \tag{2.1}$$



(a) (b)

Figure 2.3: Andamento temporale della concentrazione di alternaria e graminacee nel 2020 (zoom sul periodo estivo) con interpolazione *loess*.

Per l'unica data in cui la concentrazione non è stata rilevata in nessuna delle due stazioni (09/12/2020), il dato è stato imputato come media tra il valore del giorno precedente e quello successivo. Per l'ozono e il biossido di azoto è stato considerato il valore massimo giornaliero, rilevato presso la stazione di Mandria. Il biossido di azoto non presentava valori mancanti, mentre l'unico dato mancante per l'ozono (di data 20/10/2020) è stato imputato come media tra il valore precedente e quello successivo.

Dati 2017. Gli allergeni non sono stati rilevati dal 1° gennaio al 20 febbraio, e dal 18 al 31 dicembre. Osservando l'andamento temporale di queste sostanze (fig. 2.5), si possono imputare questi valori a 0 (come viene confermato anche guardando gli andamenti temporali nel 2020-2021). I dati sul particolato, mantenuti per coerenza solo per la stazione di Mandria, non presentavano valori mancanti. In una giornata non sono stati rilevati i valori di umidità massima e minima, e per tre giornate quelli della temperatura. Questi dati sono stati imputati come media tra il giorno precedente ed il giorno successivo.

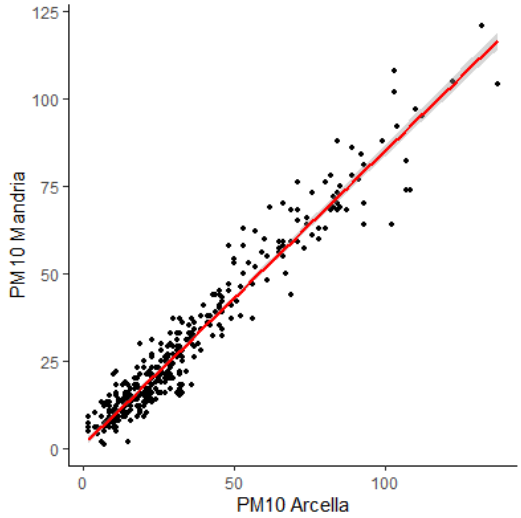


Figure 2.4: Regressione lineare tra concentrazioni di PM10 rilevate nella stazione di Arcella e Mandria (2020).

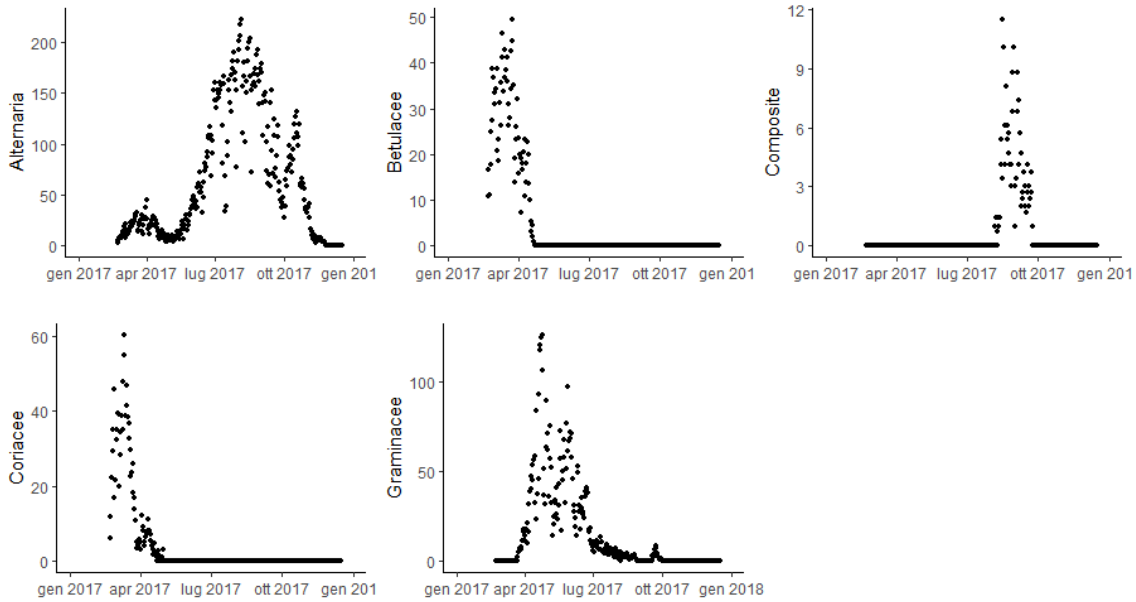


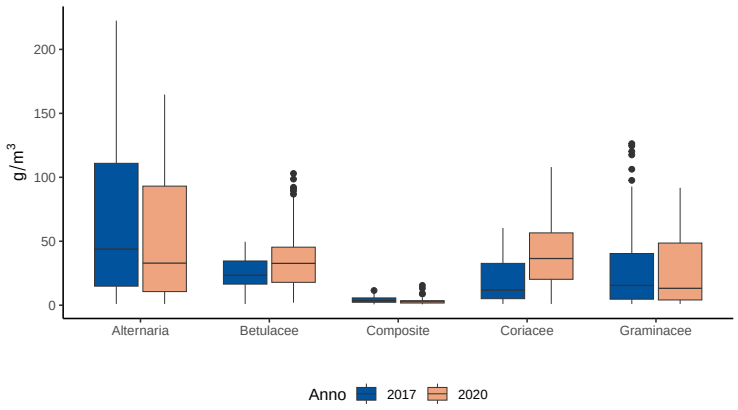
Figure 2.5: Andamento temporale della concentrazione degli allergeni nel 2017.

	2017	2020
Allergeni		
Alternaria	23 (3, 88)	15 (3, 71)
Betulacee	0 (0, 0)	0 (0, 6)
Composite	0 (0, 0)	0 (0, 0)
Coriacee	0 (0, 0)	0 (0, 19)
Graminacee	0 (0, 11)	0 (0, 12)
Clima		
Temperatura	14 (8, 22)	14 (8, 21)
Umidità minima	49 (40, 62)	52 (42, 65)
Umidità massima	94 (86, 99)	98 (92, 100)
Precipitazioni	0.0 (0.0, 0.2)	0.0 (0.0, 0.2)
Inquinanti		
Particolato (PM10)	32 (22, 55)	21 (13, 38)
Biossido di azoto (NO_2)	57 (43, 77)	40 (30, 56)
Ozono (O_3)	90 (51, 116)	89 (54, 120)

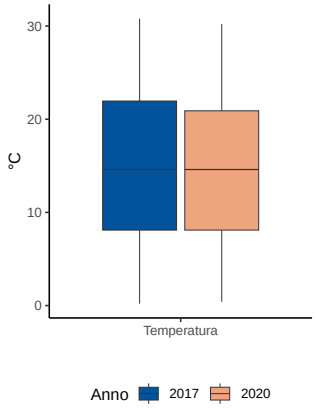
Table 2.3: Mediana (IQR) per variabili ambientali.

2.2.1 Analisi descrittive

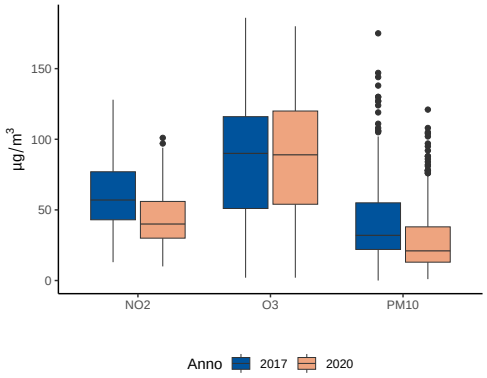
Alcune statistiche descrittive per le variabili ambientali sono riportate in tab. 2.3 sottoforma di mediana (intervallo interquartile). Per allergeni e precipitazioni, data l'alta prevalenza di giorni in cui queste variabili sono assenti, viene fornita anche la distribuzione per i valori diversi da 0 (tab. 2.4). Si nota come nel 2020 il periodo di permanenza nell'aria sia aumentato per tutti gli allergeni, ad eccezione dell'alternaria che tuttavia subisce una diminuzione della concentrazione mediana. Aumentano anche i giorni di pioggia, ma il valore mediano di millimetri caduti rimane pressoché costante rispetto al 2017. Le distribuzioni delle variabili ambientali vengono rappresentate tramite box-plot in figura 2.6 (per allergeni e precipitazioni il grafico si concentra sui valori non nulli per una migliore interpretabilità).



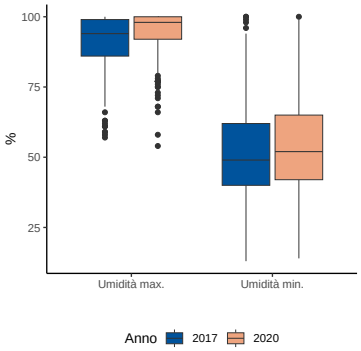
(a) Allergeni



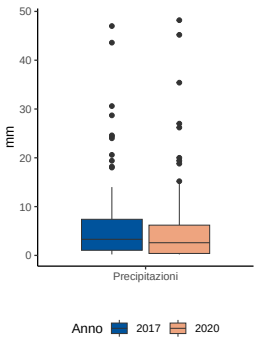
(b) Temperatura



(c) Inquinanti



(d) Umidità



(e) Precipitazioni

Figure 2.6: Box-plot delle variabili ambientali (valori non nulli per allergeni e precipitazioni).

	2017			2020		
	N	Distribuzione	% nulli	N	Distribuzione	% nulli
Alternaria	278	44 (15, 111)	24%	278	33 (11, 93)	24%
Betulacee	60	23 (16, 35)	84%	102	33 (18, 45)	72%
Composite	48	3.7 (2.3, 5.7)	87%	56	2.7 (1.7, 3.6)	85%
Coriacee	65	12 (5, 33)	82%	119	36 (20, 57)	67%
Graminacee	170	16 (5, 40)	53%	176	13 (4, 49)	52%
Precipitazioni	98	3 (1, 7)	73%	116	2.6 (0.4, 6.2)	68%

Table 2.4: Mediana (IQR) per allergeni e precipitazioni non nulli.

Nei grafici in figura 2.7 si opera un confronto fra l'andamento temporale degli allergeni nei due periodi considerati. In particolare, si osserva che le concentrazioni di composite e graminacee si mantengono sostanzialmente uguali nei due periodi. Nel 2020 i livelli di alternaria appaiono invece più bassi soprattutto nel periodo estivo, quando la concentrazione raggiunge il picco, mentre il rilascio dei pollini di betulacee e coriacee avviene in un momento dell'anno più tardivo, e con concentrazioni che raggiungono picchi più alti, rispetto a quanto registrato nel 2017. La composizione dei pollini nell'aria può infatti variare da un anno all'altro poiché determinata dalle condizioni climatiche e ambientali, in grado di influenzare i processi biologici legati alla fioritura delle piante e dei funghi [ARPAV, 2022].

Similmente, in figura 2.8 viene mostrato un confronto tra le variabili climatiche. In questo caso è stata aggiunta una curva di interpolazione, separatamente per anno, con il metodo *loess* con parametro di *span* pari a 0.5, in modo da evidenziare maggiormente eventuali differenze tra i due andamenti temporali. La variabilità dell'umidità massima sembra più contenuta nel 2020, e questo porta ad ottenere valori mediamente più alti, mentre la temperatura a gennaio 2021 sembra più alta rispetto a gennaio 2017 (possibile effetto del cambiamento climatico). Per quanto riguarda le polveri sottili (fig. 2.9), emerge che nel 2020 la concentrazione è minore, soprattutto nei mesi tra aprile e ottobre, segno di un possibile effetto del *lockdown* da pandemia Covid-19. La differenza appare più accentuata per il biossido di azoto, che nel 2020 presenta una concentrazione minore,

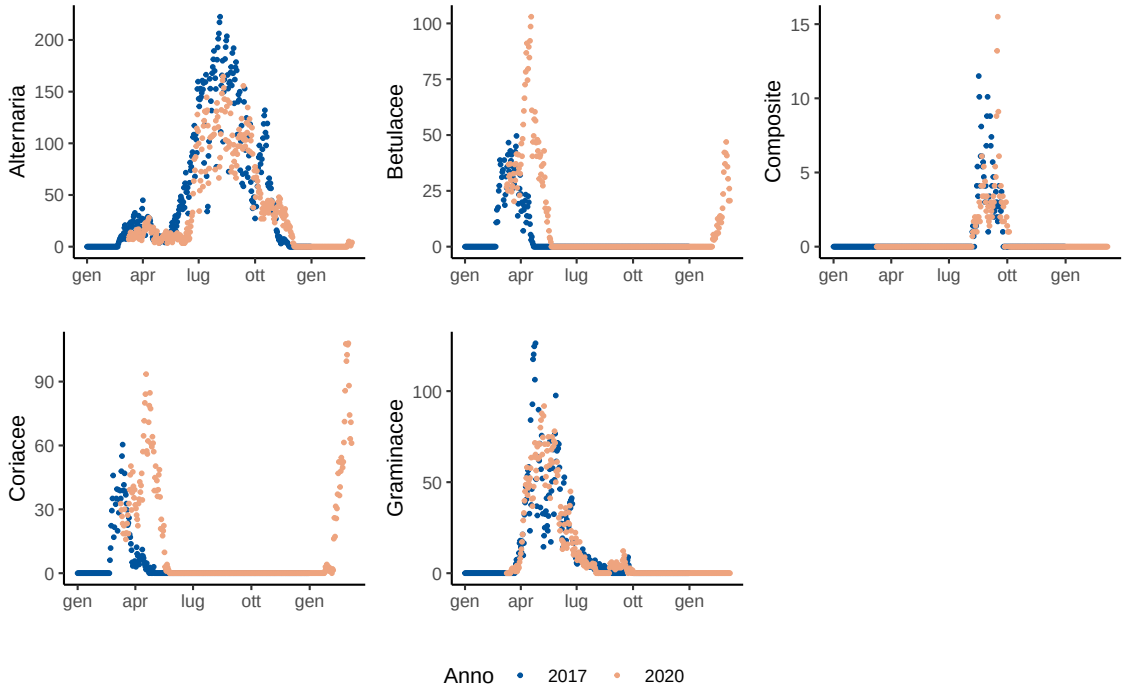


Figure 2.7: Andamento temporale della concentrazione degli allergeni nel 2017 e nel 2020.

mentre l'ozono sembra rimanere invariato, fatta eccezione a gennaio similmente a quanto accade per la temperatura. Questi risultati confermano quanto anticipato dalle statistiche descrittive marginali e dai box-plot.

In figura 2.10 vengono mostrate le correlazioni significative (p-*value* < 0.05) fra variabili ambientali nel 2017 e nel 2020 rispettivamente. Tra le correlazioni più alte, si osserva che le betulacee sono positivamente associate con le coriacee (i pollini vengono rilasciati nello stesso periodo). L'alternaria è positivamente correlata con la temperatura, infatti le spore di alternaria vengono rilasciate ad alte temperature (tra i 18 e i 32 gradi). Fra le altre correlazioni tra concentrazioni di allergeni, anche se meno rilevanti, si trovano l'alternaria con le composite e le betulacee con le graminacee (correlazione significativa solo nel 2020). Le polveri sottili risultano negativamente correlate con la temperatura e, di conseguenza, con buona parte degli allergeni, mentre risultano positivamente associate con l'umidità. Il biossido di azoto risulta positivamente correlato con il particolato, mentre l'ozono

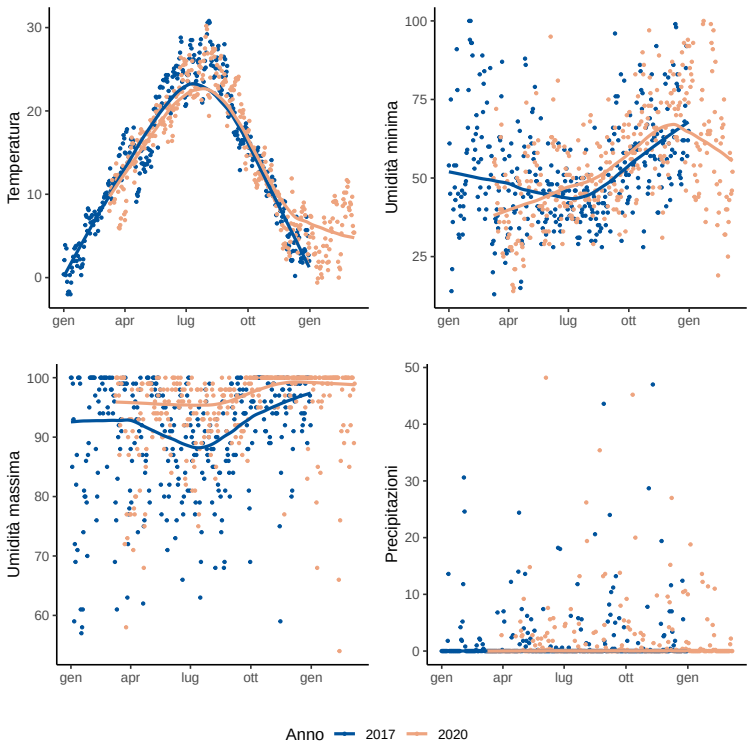


Figure 2.8: Andamento temporale delle variabili climatiche nel 2017 e nel 2020.

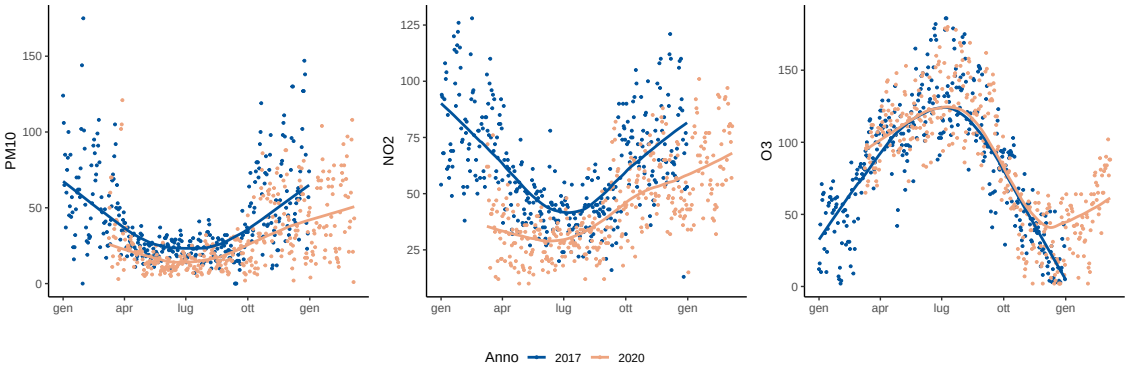


Figure 2.9: Andamento temporale degli inquinanti nel 2017 e nel 2020.

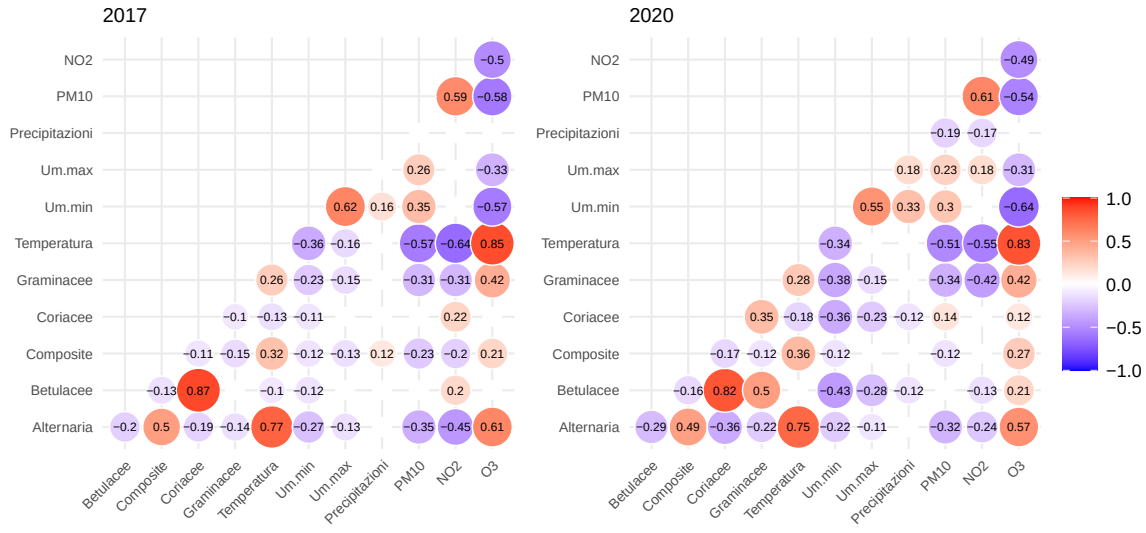


Figure 2.10: Correlazioni significative (p-*value* < 0.05) fra variabili ambientali nel 2017 e nel 2020.

correla positivamente con la temperatura. Le correlazioni trovate sono confermate dagli andamenti temporali: le spore di alternaria seguono l'andamento della temperatura, rispecchiata anche dalla concentrazione di ozono; le concentrazioni di polveri sottili e biossido di azoto aumentano entrambe nei periodi più freddi, mentre hanno una concentrazione minore nei mesi primaverili/estivi; i pollini delle composite vengono rilasciati solo tra agosto e ottobre, raggiungendo il picco nel momento in cui betulacee e coriacee non sono presenti. Alcuni diagrammi di dispersione tra le variabili maggiormente correlate sono mostrati, per i dati del 2020, in figura 2.11. Infine, in figura 3.11, si mostrano le autocorrelazioni per gli inquinanti. Nel 2020 le oscillazioni, soprattutto per il particolato, suggeriscono la presenza di stagionalità settimanale, meno evidente nel 2017.

Le analisi di modellazione che seguono nei capitoli successivi si concentreranno dapprima sui dati del 2020, per poi estendersi anche ai dati del 2017 al fine di operare un confronto fra i due periodi.

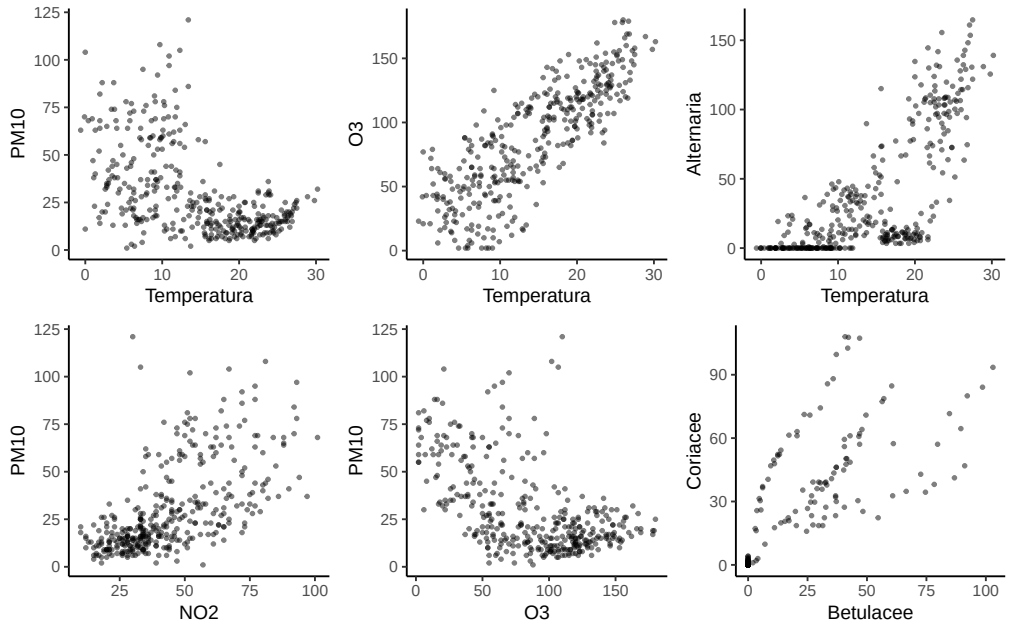
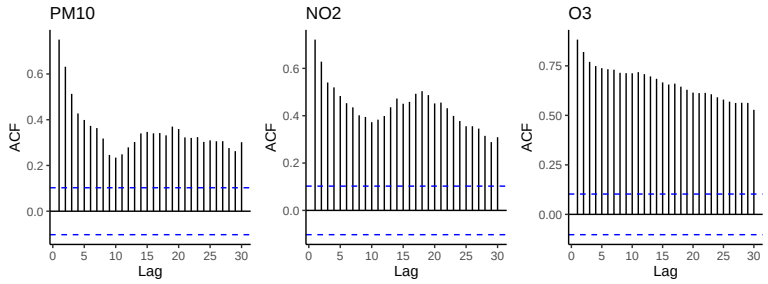
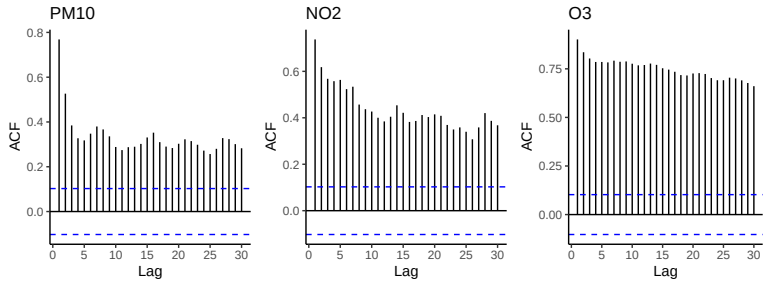


Figure 2.11: Diagrammi di dispersione delle coppie di variabili con correlazioni maggiori nel 2020.



(a) 2017



(b) 2020

Figure 2.12: Autocorrelazioni degli inquinanti nel 2017 (a) e nel 2020 (b).

Capitolo 3

Modelli per il numero di accessi giornalieri

3.1 Analisi statistiche

In questo capitolo si vuole esaminare l'influenza dei fattori ambientali sul numero di accessi giornalieri in pronto soccorso. L'unità statistica è il giorno, e ogni patologia costituisce una variabile risposta di conteggio su cui si vuole studiare l'impatto delle variabili ambientali. Per ogni giornata, sono stati dunque contati il numero di accessi totali per patologie legate alle alte vie respiratorie e per ciascuna causa specifica al pronto soccorso (oltre che gli accessi per trauma). In seguito, ci si riferirà agli accessi totali per patologie legate alle Alte Vie Respiratorie con l'abbreviazione di 'Accessi AVR'. Come visto dalla tabella 2.3, le diagnosi di tracheite sono in numero molto ridotto, pertanto questa patologia non verrà considerata singolarmente ma all'interno di una macrodiagnosi che comprende patologie legate all'apparato respiratorio (laringite, faringite, tonsillite, tracheite), oltre che ad essere considerata all'interno degli accessi AVR.

3.1.1 *Splines* di regressione

Il termine *spline* viene utilizzato per esprimere una funzione polinomiale a tratti per lo studio della relazione tra una risposta y e una covariata x. Scelti K punti sull'asse delle x, detti nodi, la funzione $f(x)$ viene costruita imponendo il vincolo di interpolazione (la funzione deve passare per il valore osservato della y in ciascun nodo) e di continuità nei nodi per evitare la presenza di punti angolosi. Tra due nodi, la funzione coincide con un polinomio che, se di ordine 3, dà origine alla cosiddetta *spline* cubica. Se si aggiunge il vicolo che i due polinomi alle estremità siano linee rette, la funzione risultante viene detta *spline* cubica naturale.

La *spline* di regressione, che non richiede il vincolo di interpolazione, è un approccio parametrico per lo studio della relazione tra una risposta y e una covariata x considerando un modello del tipo:

$$y_i = f(x_i) + \epsilon_i \tag{3.1}$$

dove $\epsilon_i \sim N(0, \sigma^2)$ e f è una funzione di lisciamento espressa come una combinazione lineare di funzioni di base b_j di dimensione q:

$$f(x) = \sum_{j=1}^{q} \beta_j b_j(x) \tag{3.2}$$

La *spline* è detta di regressione perché, una volta definiti i nodi e le funzioni di base, i parametri β possono essere stimati tramite il metodo dei minimi quadrati (minimizzazione della devianza residua). Un modo di controllare il lisciamento della funzione è agire sul numero di funzioni di base, dei nodi e della loro posizione. Un metodo alternativo, da cui si ottiene la *spline* di lisciamento, consiste invece nel considerare il criterio dei minimi quadrati

$$D(f, \lambda) = \sum_{i=1}^{n} [y_i - f(x_i)]^2 \tag{3.3}$$

nella sua versione penalizzata:

$$D(f, \lambda) = \sum_{i=1}^{n} [y_i - f(x_i)]^2 + \lambda \int_{-\infty}^{\infty} [f''(t)]^2 dt \tag{3.4}$$

dove λ è detto parametro di lisciamento. La derivata seconda è una misura dell'irregolarità della curva: porre $\lambda = 0$ equivale quindi a non applicare nessuna penalizzazione all'irregolarità, mentre una penalizzazione massima ($\lambda \to \infty$) riconduce alla retta di regressione ai minimi quadrati. La funzione f che risolve il problema di minimizzazione della 3.4 è detta *spline* di lisciamento. In particolare, si dimostra che la f ottimale è una *spline* cubica naturale con un nodo per ogni valore osservato della x [Green and Silverman, 1994].

3.1.2 Modelli additivi generalizzati

I modelli additivi costituiscono un valido approccio di modellazione semi-parametrica grazie all'ampia flessibilità e la facilità di interpretazione, e permettono di imporre ai dati una struttura semplice e con poche assunzioni.

Sia $Y = (Y_1, Y_2, \ldots, Y_T)$ il vettore casuale della variabile risposta, α un termine costante che rappresenta l'intercetta del modello, $X_j (j = 1, \ldots, p)$ variabili esplicative osservate, $f_j (j = 1, \ldots, p)$ delle funzioni di lisciamento univariate, ed ϵ il termine di errore. La formulazione generale del modello additivo risulta

$$Y = \alpha + \sum_{j=1}^{p} f_j(X_j) + \epsilon \tag{3.5}$$

e si basa sulle seguenti assunzioni:

- gli errori ϵ sono indipendenti dalle variabili esplicative X_j
- $E(\epsilon) = 0$
- $Var(\epsilon) = \sigma^2 I$
- le funzioni f_j devono avere un andamento sufficientemente regolare e liscio.

Per garantire l'identificabilità del modello, si assume inoltre che

$$E[f_j(X_j)] = 0 \qquad \forall j = 1, \ldots, p \tag{3.6}$$

ovvero che le funzioni f_j siano centrate in 0. La struttura del modello additivo è dunque analoga a quella del modello lineare classico, ma coinvolge funzioni che permettono di rilassare il vincolo di linearità tra le esplicative e la risposta. Le funzioni f_j possono essere dei lisciatori non parametrici quali *splines*, *loess*, ecc., ma possono altresì essere funzioni parametriche ed esprimersi ad esempio come $f_j(X_j) = \beta_j X_j$. Chiaramente, una formulazione di questo tipo per tutte le funzioni di base introdotte nel modello riporta alla regressione lineare classica. Spesso la scelta del lisciatore non è cruciale, e possono essere scelti metodi diversi per diverse funzioni f_j.

Per stimare un modello additivo si può usare l'algoritmo di *backfitting* (anche detto algoritmo di Gauss-Seidel) proposto da Hastie and Tibshirani [1990]. Se però ci si limita a lisciatori che siano *spline* di lisciamento, la stima dei parametri avviene mediante massimizzazione della verosimiglianza penalizzata con selezione automatica del parametro di lisciamento tramite il metodo della massima verosimiglianza ristretta (REML, *Restricted Maximum Likelihood*), come proposto da Wood [2017].

In questo studio, la scelta delle funzioni di lisciamento ricade sulla *spline* di lisciamento, che nel caso bivariato viene chiamata *thin-plate spline*. La formulazione 3.5 può essere infatti ampliata con l'aggiunta di termini di interazione, rappresentati da funzioni che coinvolgono più di una variabile. Ad esempio, il modello additivo che comprende tutte le interazioni doppie risulta:

$$Y = \alpha + \sum_{j=1}^{p} f_j(X_j) + \sum_{j=1}^{p} \sum_{k<j} f_{kj}(X_k, X_j) + \epsilon \tag{3.7}$$

Il modello additivo può essere generalizzato in maniera simile a quanto avviene per il modello lineare:

$$g(\mu) = \alpha + \sum_{j=1}^{p} f_j(X_j) \tag{3.8}$$

dove $g(\cdot)$ è la funzione di legame e μ è la media della variabile risposta condizionata alle esplicative, ossia $\mu = E(Y|X_1, \ldots, X_p)$. Nel caso specifico di questo capitolo, vengono adattati modelli additivi generalizzati (*Generalized Additive Model*, GAM) per variabili di conteggio. La funzione di legame utilizzata è quindi il logaritmo, così come avviene per il modello di Poisson con legame canonico, che si formula come:

$$\log(\mu) = \alpha + \sum_{j=1}^{p} \beta_j X_j \tag{3.9}$$

3.1.3 Stima del modello GAM

L'idea di base per la stima del modello GAM è ricondurre quest'ultimo alla formulazione di un GLM e applicare l'algoritmo dei minimi quadrati pesati iterati (*Iteratively Reweighted Least Squares*, IRLS) alla verosimiglianza penalizzata, utilizzando quello che viene chiamato algoritmo PIRLS (*Penalized*-IRLS) [Wood, 2017].

Si consideri il caso di un modello GAM generico:

$$g(\mu) = f_1(x_1) + f_2(x_2) + f_3(x_3) + \dots \tag{3.10}$$

dove $\mu = E(Y)$, con Y appartenente a una famiglia esponenziale. Ciascuna funzione f_j può essere espressa nella forma

$$f_j(x_j) = \sum_{i=1}^{q_j} \beta_{ji} b_{ji}(x_j) \tag{3.11}$$

Ponendo

- $\mathbf{f_j} = [f_j(x_{j1}), f_j(x_{j2}), \dots, f_j(x_{jn})]^T$
- $\beta_j = [\beta_{j1}, \beta_{j2}, \dots, \beta_{jq_j}]^T$
- $\mathbf{X_j}$ matrice $n \times q_j$ in cui $\mathbf{X_{j,ik}} = b_{jk}(x_j i)$

si ha

$$\mathbf{f_j} = \mathbf{X_j}\beta_j \tag{3.12}$$

e il modello può essere riscritto come

$$g(\mu) = X\beta \tag{3.13}$$

in cui $X = [\mathbf{X_1} : \mathbf{X_2} : \dots]$ e $\beta = [\beta_1^T, \beta_2^T, \dots]^T$. Il modello così definito ha la formulazione di

un GLM, con generica funzione di verosimiglianza $l(\beta)$. La funzione di verosimiglianza penalizzata risulta:

$$l_p(\beta) = l(\beta) - \frac{1}{2}\sum_j \lambda_j \beta^T S_j \beta = l(\beta) - \frac{1}{2}\beta^T S \beta \tag{3.14}$$

dove S_j è una matrice di parametri noti, $S = \sum_j \lambda_j S_j$, e il termine di penalizzazione è governato dai parametri di lisciamento λ_j. La penalità $\beta^T S \beta$ è quindi funzione di una matrice di penalità S e dei coefficienti del modello β, che costituiscono i pesi delle funzioni di base. La penalità agisce comprimendo le stime di β (effettua uno *shrinkage*), in modo da attribuire minor peso alle funzioni più ondulate.

Considerando per il momento noti i parametri di lisciamento, l'algoritmo PIRLS procede iterativamente fino a convergenza: alla $(k+1)$-esima iterazione dell'algoritmo viene stimato il vettore di coefficienti

$$\hat{\beta}^{(k+1)} = (X^T W^{(k)} X + S)^{-1} X^T W^{(k)} z^{(k)} \tag{3.15}$$

utilizzando come variabile risposta la variabile dipendente aggiustata $z^{(k)}$ ottenuta nell'iterazione precedente e con generica componente

$$z_i^{(k)} = \mathbf{x}_i \hat{\beta}^{(k)} + (y_i - \mu_i) g'(\mu_i) \tag{3.16}$$

La matrice diagonale di pesi $W^{(k)}$, anch'essa calcolata all'iterazione k-esima, ha il generico elemento

$$w_i^{(k)} = \frac{1}{V(\mu_i^{(k)}) g'(\mu_i^{(k)})^2} \tag{3.17}$$

Per la scelta dei parametri di lisciamento viene usato il criterio REML, che può essere visto in maniera semplice e intuitiva attraverso un approccio bayesiano in cui viene specificata una distribuzione a priori sui coefficienti β del tipo:

$$\beta \sim N(0, \sigma^2 S^- / \lambda) \tag{3.18}$$

dove S^- indica la pseudo-inversa di S. L'interpretazione bayesiana dei parametri di liscia-

mento dà al modello la struttura di un modello a effetti misti. I parametri di lisciamento vengono considerati come effetti casuali di un modello ad effetti misti e stimati massimizzando il criterio REML:

$$l_r(\hat{\beta}, \lambda) = log \int f(y|\beta) f(\beta) d\beta \tag{3.19}$$

Il criterio viene risolto tramite i seguenti passaggi, ripetuti fino a convergenza:

1. Dato un valore iniziale di λ, β viene stimato con l'algoritmo PIRLS;
2. λ viene aggiornato massimizzando la verosimiglianza ristretta.

3.1.4 Effetti ritardati

Le sostanze presenti nell'aria, una volta inalate, possono avere un effetto sulla salute immediato ma anche che si prolunga per più giorni. Per cogliere questo effetto è possibile inserire nel modello le variabili esplicative ritardate di alcuni giorni, oppure l'effetto cumulato di più giorni. O ancora, si può inserire una media mobile che coglie l'effetto medio ritardato su più giorni (questo non è altro che l'effetto cumulato ritardato riscalato per una costante, ma serve ad avere un valore composito espresso nella stessa misura di grandezza delle misurazioni giornaliere). In questo studio, si ritiene necessario considerare l'effetto ritardato degli inquinanti fino a 3 giorni rispetto al giorno di riferimento, inserendo anche il valore medio con i 3 valori ritardati. Dato X_t l'inquinante X al giorno t, verranno quindi definiti X_{t-1}, X_{t-2}, X_{t-3} e $\frac{1}{4}(X_t + X_{t-1} + X_{t-2} + X_{t-3})$.

3.1.5 Confondimento

Un fattore di confondimento è una variabile in grado di generare un'associazione apparente oppure di mascherarne una realmente esistente tra una certa esposizione e un esito di salute di interesse. Affinché una variabile possa considerarsi confondente deve soddisfare tre caratteristiche: 1) deve essere un fattore di rischio/protezione per l'esito di interesse; 2) deve essere associata, in modo causale o non causale, con l'esposizione; 3) non deve essere causata né dall'esposizione né dall'esito di salute. Il confondimento è un fenomeno di distorsione che genera la cosiddetta relazione spuria, e

che va pertanto controllato. Per tenere conto del confondimento a posteriori, cioè dopo la raccolta dei dati, si possono adottare diverse tecniche quali l'analisi stratificata, l'appaiamento e l'analisi multivariata.

Nel caso specifico di questo studio, ci sono alcune caratteristiche che possono distorcere la stima dell'effetto che allergeni e inquinanti hanno sugli accessi in pronto soccorso. Si può parlare ad esempio delle variabili meteorologiche, che possono agire sulla presenza nell'aria di alcuni aeroallergeni e sostanze inquinanti, oltre che ad avere un'influenza sulla salute. I giorni della settimana, distinguendo in particolare i giorni feriali da quelli festivi, possono agire sul livello di inquinamento provocato dal traffico urbano e dalle attività industriali, ed avere al contempo un'influenza sull'accesso al pronto soccorso e la gestione di quest'ultimo.

Un discorso a parte va fatto per le restrizioni dovute alla pandemia. In maniera più o meno restrittiva, le misure di contenimento adottate durante l'emergenza pandemica limitavano gli spostamenti e aumentavano il senso di paura e disagio nell'uscire di casa e frequentare luoghi a rischio di infezione come il pronto soccorso. La pandemia ha inoltre comportato una diversa gestione del pronto soccorso da parte delle strutture ospedaliere, che poteva subire cambiamenti anche al variare delle fasi di *lockdown* e dell'incidenza nella popolazione dell'infezione da Covid-19. In Italia, tra marzo 2020 e marzo 2021, si sono susseguite cinque fasi che stabilivano accorgimenti e misure di contenimento più o meno pesanti a seconda dell'entità della pandemia e degli sviluppi sulla gestione della stessa. Le restrizioni, intese come tutto ciò che poteva limitare l'accesso in pronto soccorso e l'esposizione alle sostanze nell'aria, esercitano dunque un'influenza, variabile nel tempo, che è necessario quantificare e controllare. Con questo obiettivo, gli accessi per trauma possono essere d'aiuto, poiché non sono influenzati da variabili ambientali. Ovvero, se a parità delle altre condizioni c'è una differenza tra gli accessi per trauma del 2020 e quelli del 2017, è plausibile pensare che questa differenza possa essere attribuita alle restrizioni da pandemia, quantificate dalla differenza percentuale tra gli accessi per trauma nei due anni. Una differenza molto alta in un certo periodo (ad esempio, in un dato mese), indica che in quel periodo c'è stata una forte influenza della pandemia in termini di limitazioni, mentre una differenza blanda indica che la pandemia esercitava meno pressione.

3.2 Risultati

3.2.1 Analisi esplorative

In figura 3.1 vengono riportate le distribuzioni sul numero di accessi per ogni patologia considerata, compresi gli accessi totali per patologia, gli accessi per cause legate all'apparato respiratorio e i traumi. Tutte le distribuzioni risultano asimmetriche e la maggior parte delle diagnosi presenta una grande quantità di zeri, segno che in molte giornate non sono stati registrati accessi per quella causa. In 62 giornate non sono stati registrati accessi totali AVR, e in 10 giornate non sono stati registrati accessi né AVR né per trauma. In figura 3.2 vengono mostrati gli andamenti temporali degli accessi in pronto soccorso. Si nota che gli accessi subiscono un innalzamento nel periodo estivo, anche se di debole densità. Il picco è visibile soprattutto per le patologie con accessi più numerosi, come la faringite e le macrodiagnosi. In tabella 3.1 sono riportate media e varianza per ciascuna variabile di conteggio. Per le patologie con eccesso di zeri si osserva che tipicamente la varianza è superiore alla media, suggerendo la presenza di sovradispersione. Dall'analisi delle correlazioni tra il numero di accessi per le varie patologie (fig. 3.3), non si riscontrano correlazioni significative, a meno di quelle tra le macrodiagnosi e le rispettive componenti. Sono presenti autocorrelazioni significative, ma appaiono sporadiche e non persistenti (fig. 3.4).

Ai fini delle analisi successive, si vuole indagare se l'eccesso di zeri e l'eventuale sovradispersione siano elementi da tenere in considerazione. A questo scopo, per ogni variabile risposta sono stati adattati, con la sola intercetta, il modello di Poisson, il modello binomiale negativo, il modello di Poisson con inflazione di zeri, il modello binomiale negativo con inflazione di zeri (per completezza è stato adattato, anche se poco idoneo al tipo di variabile risposta, anche il modello gaussiano). Per verificare quale distribuzione si adattasse meglio ai dati marginali, i modelli vengono messi a confronto tramite l'indice AIC (tab. 3.2), osservando che generalmente l'utilizzo di metodi che tengano conto della sovradispersione e/o dell'eccesso di zeri non apporta notevoli miglioramenti all'adattamento rispetto al modello di Poisson semplice.

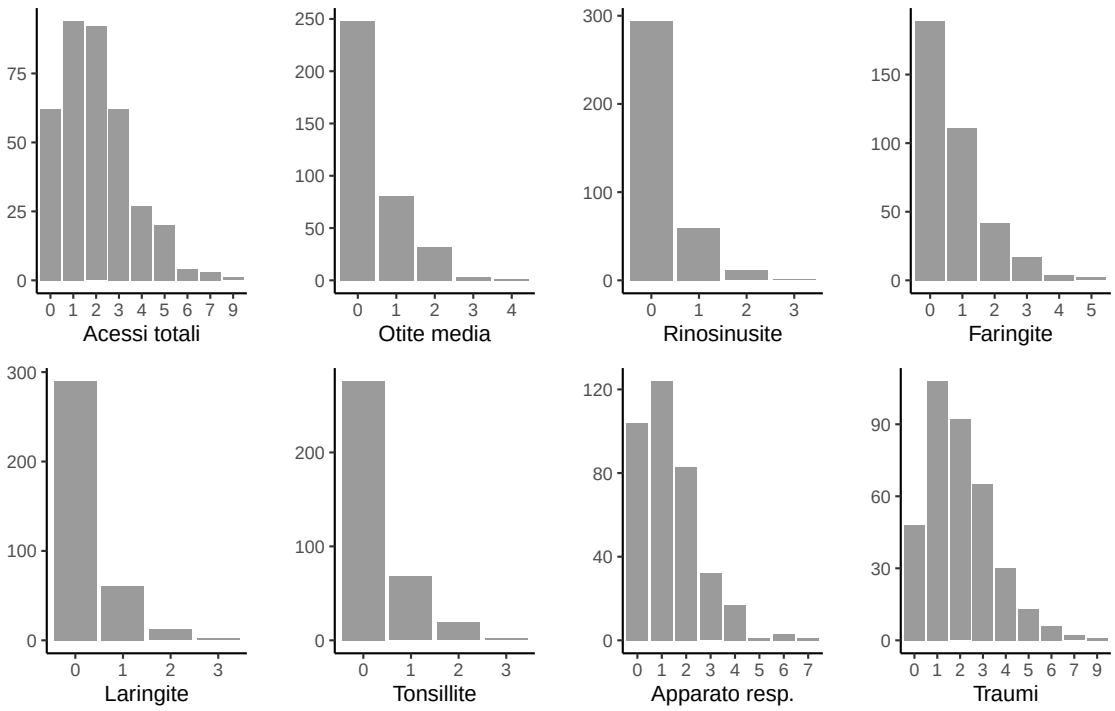


Figure 3.1: Distribuzione assoluta del numero di accessi giornaliero per le patologie considerate.

	Media	Varianza
Accessi AVR	1.99	1.55
Otite media	0.43	0.71
Rinosinusite	0.23	0.50
Faringite	0.75	0.97
Laringite	0.25	0.54
Tonsillite	0.31	0.59
Apparato respiratorio	1.33	1.24
Traumi	2.00	1.48

Table 3.1: Media e varianza del numero di accessi per ciascuna patologia.

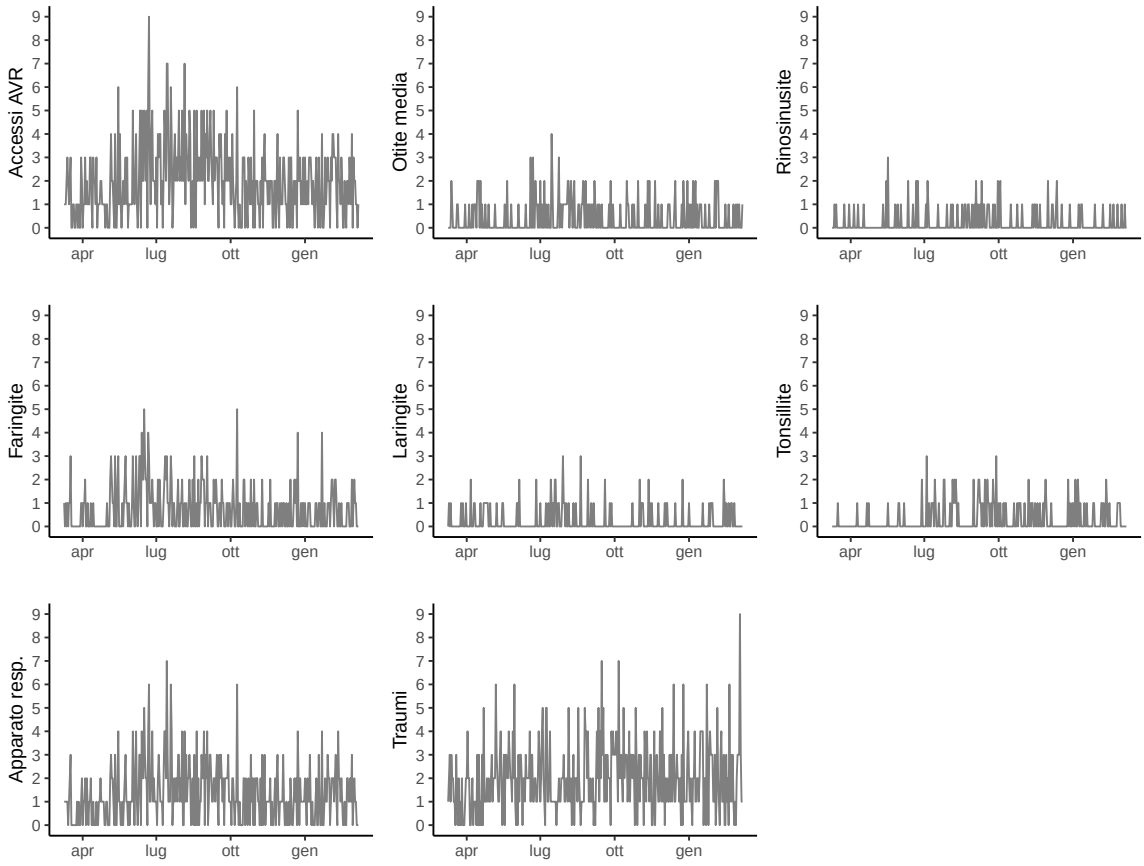


Figure 3.2: Andamento temporale del numero di accessi giornalieri.

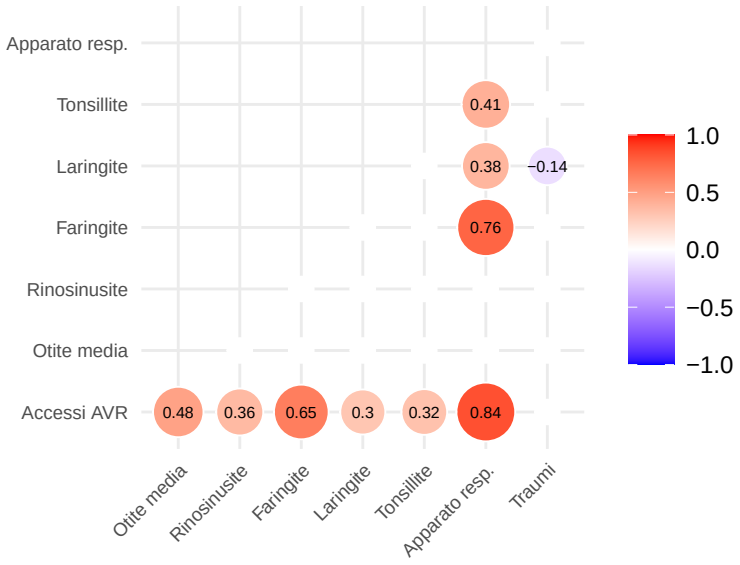


Figure 3.3: Correlazioni significative (p-*value* < 0.05) fra numero di accessi.

	Gaussiano	Poisson	Binomiale negativa	Poisson Zero-Inflated	Bin.negativa Zero-Inflated
Accessi AVR	1361	1298	1293	1295	1295
Otite media	789	644	641	639	641
Rinosinusite	539	436	436	435	437
Faringite	1017	870	862	863	864
Laringite	583	461	459	459	461
Tonsillite	656	524	523	521	523
Apparato resp.	1194	1101	1099	1101	1101
Traumi	1325	1263	1264	1265	1266

Table 3.2: AIC modelli con sola intercetta.

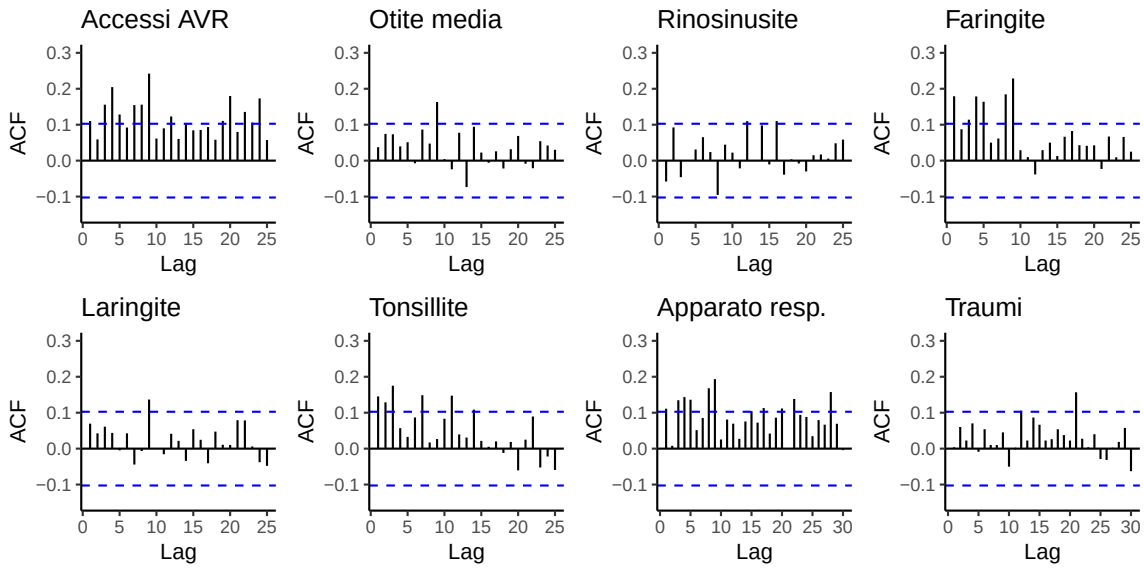


Figure 3.4: Autocorrelazioni per il numero di accessi.

3.2.2 Effetti univariati

Come prima misura di effetto delle variabili ambientali sul numero di accessi in pronto soccorso, sono stati adattati modelli univariati di regressione di Poisson con legame canonico. I risultati, espressi in termini di rapporto fra tassi di incidenza, sono esposti in tab. 3.3, e sono associati ad un incremento di 10 unità della variabile esplicativa. A titolo esemplificativo, si interpreta l'effetto dell'alternaria sugli accessi AVR: l'incremento di 10 spore/m^3 della concentrazione di altenaria comporta un aumento del 5% del numero medio di accessi in pronto soccorso per cause legate alle alte vie respiratorie. Per la maggior parte delle patologie e per le macrodiagnosi (accessi AVR e apparato respiratorio) la temperatura presenta un effetto positivo sul numero medio di accessi. Di conseguenza, dipendendo direttamente dalla temperatura, anche l'alternaria (e in alcuni casi le composite) e l'ozono risultano essere fattori di rischio. Al contrario, PM10 e NO_2, negativamente associati con la temperatura, presentano un effetto negativo, associato a una diminuzione del numero medio di accessi all'aumentare della concentrazione.

Per cogliere eventuali effetti non lineari, mantenendo comunque una buona interpretabilità dei risultati, sono stati adattati modelli GAM (*Generalized Additive Model*) di tipo Poisson, anch'essi

	Accessi AVR	Apparato resp.	Traumi
Alternaria	**1.052** ***	**1.048** ***	1.001
Betulacee	**0.921** **	**0.886** ***	**0.933** ***
Composite	**1.535** *	1.413	**1.657** **
Coriacee	**0.930** ***	**0.920** ***	**0.966** *
Graminacee	0.976	0.980	0.989
PM10	**0.949** **	**0.948** *	1.013
O_3	**1.035** ***	**1.036** ***	0.990
NO_2	**0.953** **	0.954	**1.050** **
Temperatura	**1.362** ***	**1.377** ***	0.993
Umidità min.	0.992	0.992	1.029
Umidità max.	0.971	0.956	**1.136** *
Precipitazioni	0.923	0.814	0.991

	Otite media	Rinosinusite	Faringite	Laringite	Tonsillite
Alternaria	**1.065** ***	1.045	**1.029** *	**1.052** *	**1.095** ***
Betulacee	1.012	0.902	**0.859** ***	1.048	**0.750** **
Composite	1.286	**2.741** *	1.259	0.371	**2.876** **
Coriacee	0.967	0.913	**0.890** ***	1.042	**0.857** **
Graminacee	0.959	0.981	1.025	1.009	**0.787** ***
PM10	0.961	0.927	**0.930** *	0.999	0.960
O_3	1.035	1.031	**1.048** **	1.042	1.002
NO_2	0.928	0.997	**0.927** *	0.906	1.054
Temperatura	**1.332** **	**1.334** *	**1.472** ***	1.294	1.246
Umidità min.	1.015	0.948	1.014	0.963	0.975
Umidità max.	1.027	0.963	1.144	**0.726** **	0.848
Precipitazioni	1.198	0.825	0.901	0.740	0.607

Table 3.3: Rischi relativi ottenuti dai modelli di Poisson univariati per la regressione del numero di accessi sulle variabili ambientali. Risultati associati a un incremento di 10 unità della variabile esplicativa. Gli asterischi indicano la significatività: *** per $p.value < 0.001$, ** per $0.001 \leq p.value < 0.01$, * per $0.01 \leq p.value < 0.05$.

univariati e utilizzando come lisciatore la spline di lisciamento (*thin-plate regression spline*). La tabella 3.4 riporta i *p-value* associati a ciascun effetto non parametrico mostrando, in termini di variabili con effetto significativo, un'analogia con il modello di Poisson. In figura 3.5 vengono rappresentati i grafici degli effetti significativi ($p\text{-}value < 0.05$). Dai risultati univariati emerge che l'alternaria ha un effetto positivo e lineare sugli accessi per l'otite media e per le patologie che riguardano l'apparato respiratorio, mentre l'effetto delle coriacee ha un andamento prima descrescente e poi crescente a partire da circa 60 grammi/m^3 (30 g/m^3 per l'otite). Le betulacee hanno un effetto che diventa crescente dopo circa 40 g/m^3 su otite e apparato respiratorio; per quest'ultimo, l'effetto delle graminacee risulta crescente a partire da circa 50 g/m^3. Le composite emergono come fattore di rischio per la rinosinusite, tonsillite e per gli accessi AVR fino a circa 5 g/m^3 (per concentrazioni maggiori la ridotta numerosità rende imprecisa la stima dell'effetto). Per quanto riguarda gli inquinanti, emerge un effetto positivo dell'ozono su accessi totali, accessi per apparato respiratorio e faringite, mentre il biossido di azoto ha un andamento pressoché costante. Le polveri sottili riportano un effetto negativo su apparato respiratorio e faringite. Data la presenza di effetti non lineari, si è deciso di concentrare le analisi successive sui modelli GAM.

3.2.3 Effetti ritardati e aggiustamento

Per indagare in modo più approfondito l'effetto degli inquinanti, sono stati adattati modelli GAM su ciascun tipo di accesso e separatamente per ogni inquinante, considerando anche gli effetti ritardati e l'aggiustamento per i possibili confondenti. Il modello aveva la seguente formulazione:

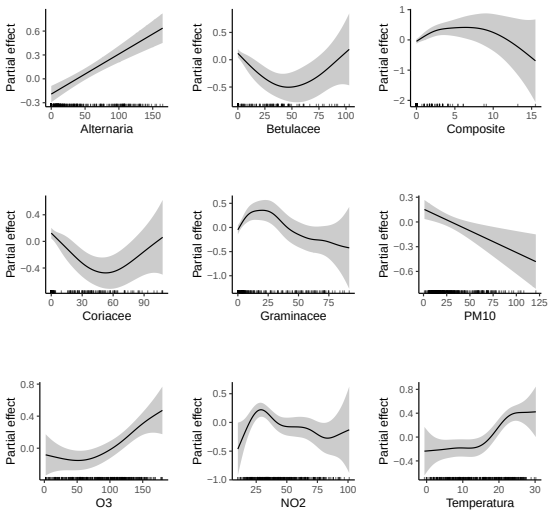
$$\log(\mu) = \alpha + f_1(X) + f_2(X_{-1}) + f_3(X_{-2}) + f_4(X_M) + \beta \cdot Wday + \gamma \cdot Festivo + f_5(Restrizioni) \quad (3.20)$$

dove

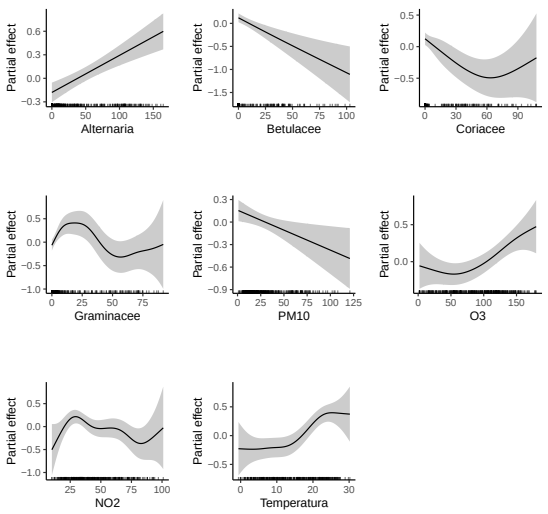
- μ è la media, condizionata alle variabili esplicative, della variabile risposta di conteggio;
- X è il valore giornaliero dell'inquinante;
- X_{-l} è il valore dell'inquinante al lag l;

	Accessi AVR	Otite media	Rinos.	Faring.	Laring.	Tons.	Appar. resp.	Traumi
Alternaria	**<0.001**	**<0.001**	0.108	**0.024**	**0.018**	**<0.001**	**<0.001**	0.147
Betulacee	**<0.001**	**0.001**	0.275	**<0.001**	0.314	**0.003**	**<0.001**	**0.001**
Composite	**0.002**	0.636	**0.004**	0.219	0.295	**0.003**	0.081	**0.006**
Coriacee	**<0.001**	**0.042**	0.109	**0.002**	0.300	**0.003**	**0.001**	0.057
Graminacee	**0.001**	0.377	0.468	**<0.001**	0.840	**0.017**	**0.001**	0.531
PM10	**0.003**	0.407	0.155	**0.014**	0.992	0.354	**0.013**	0.401
O_3	**<0.001**	0.075	0.233	**0.004**	0.072	0.943	**0.001**	0.245
NO_2	**0.003**	0.083	0.960	0.055	0.184	0.265	**0.013**	**0.008**
Temperatura	**<0.001**	**0.007**	**0.047**	**<0.001**	0.113	0.113	**<0.001**	0.889
Umidità min.	0.685	0.749	0.316	0.152	0.533	0.640	0.508	0.096
Umidità max.	**0.046**	**0.034**	0.116	0.113	**0.004**	0.149	0.461	0.083
Precipitazioni	0.443	0.252	0.465	0.086	0.522	0.113	0.541	0.898

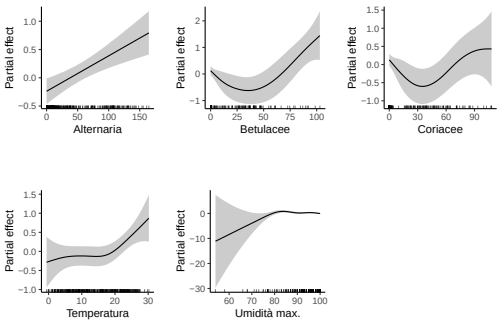
Table 3.4: P-value degli effetti ottenuti dai modelli GAM univariati.



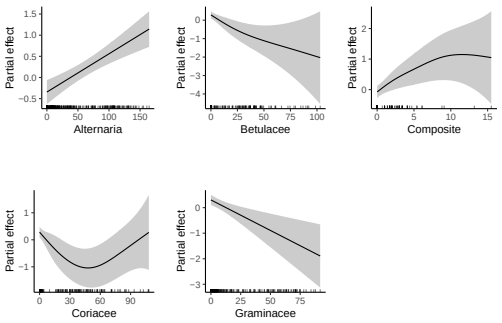
(a) Accessi AVR



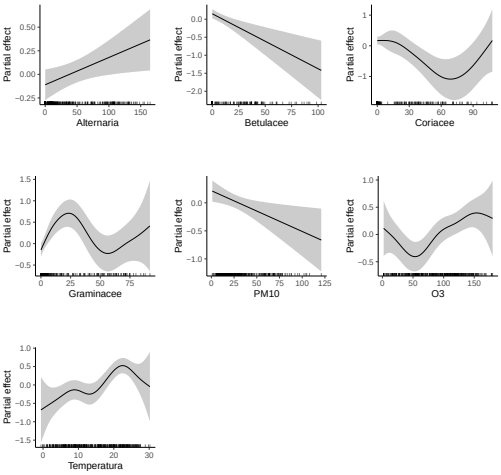
(b) Apparato respiratorio



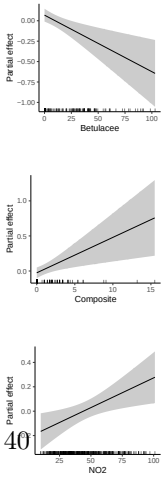
(c) Otite media



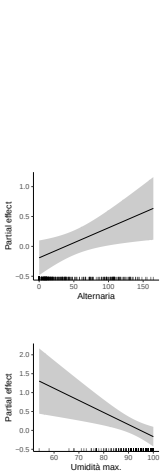
(d) Tonsillite



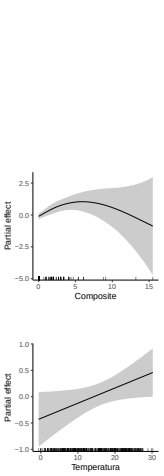
(e) Faringite



(f) Traumi



(g) Laringite



(h) Rinosinus.

Figure 3.5: Effetti significativi ottenuti dai modelli GAM univariati per ciascuna diagnosi.

- X_M è la media mobile sul valore giornaliero dell'inquinante e i 3 giorni precedenti (l'inquinante ritardato al terzo giorno, X_{-3}, non è stato inserito per evitare dipendenza lineare fra le variabili);
- f_j sono le funzioni di lisciamento (*thin-plate spline*);
- $Wday$ è il giorno della settimana, da lunedì (1) a domenica (7), mantenuta come variabile quantitativa discreta;
- $Festivo$ è una variabile dicotomica che assume valore 1 se il giorno è un sabato, una domenica o un giorno di festa, 0 altrimenti;
- *Restrizioni* è una variabile che quantifica l'impatto delle restrizioni dovute alla pandemia, definita come differenza percentuale, calcolata a livello mensile, tra gli accessi per trauma del 2017 e quelli del 2020 (dunque è una variabile costruita indipendentemente dalla risposta e dalle variabili esplicative introdotte nel modello). La forma funzionale per questa variabile è stata resa più liscia abbassando il numero di funzioni di base. Ciò permette di evitare che la forma funzionale per questa variabile colga oscillazioni irregolari dovute alla sua natura mensile.

Viene riscontrato un effetto significativo del valore medio cumulato delle polveri sottili per alcuni tipi di accessi, in particolare per gli accessi totali, gli accessi per apparato respiratorio e faringite. Il ruolo del biossido di azoto e dell'ozono risulta poco chiaro, dal momento che vengono rilevati degli effetti ritardati negativi per alcuni tipi di accessi tra cui anche gli accessi per trauma.

In uno step successivo sono state aggiunte le variabili climatiche come ulteriore aggiustamento, anch'esse in forma non parametrica. Per cercare di alleggerire il modello, poiché dalle analisi univariate le precipitazioni non risultavano significativamente associate con alcuna variabile di conteggio, questa variabile esplicativa non è stata considerata. Si riscontra un effetto significativo del valore cumulato delle polveri sottili sugli accessi per faringite: l'effetto presenta un andamento crescente a partire da circa 30 µg/m^3. Si trova inoltre un effetto positivo e lineare dell'inquinante ritardato al primo giorno sulla rinosinusite. L'ozono e il biossido di azoto continuano a presentare effetti

ritardati negativi significativi su faringite, accessi AVR e traumi (fig. 3.6).

In seguito, anche lo studio dell'effetto degli allergeni è stato approfondito inserendo l'aggiustamento per le variabili climatiche e confondenti, con un modello del tipo:

$$\begin{aligned}\log(\mu) =& \alpha + f_1(X) + f_2(Temperatura) + f_3(Umidita\ min) + f_4(Umidita\ max) + \\ & \beta \cdot Wday + \gamma \cdot Festivo + f_5(Restrizioni)\end{aligned} \tag{3.21}$$

dove X è il valore giornaliero dell'allergene. Un effetto positivo dell'alternaria si riscontra sugli accessi per tonsillite e otite media, confermando quanto trovato dalle analisi univariate. L'otite media risulta influenzata positivamente anche dalla concentrazione di betulacee, così come gli accessi AVR. Le graminacee presentano sulla faringite un effetto positivo fino a circa 25 g/m3. Si riscontra un effetto positivo delle composite sugli accessi per trauma (fig. 3.7).

Infine, per valutare congiuntamente l'effetto di inquinanti e allergeni, per ciascuna patologia è stato adattato un modello GAM con selezione stepwise in avanti basata sul criterio AIC, considerando anche le variabili ritardate degli inquinanti e il confondimento per le restrizioni. La procedura di selezione si rende necessaria dal momento che, data la bassa numerosità del campione (365), risultava poco efficiente inserire troppe covariate nel modello. Il codice utilizzato per implementare la procedura stepwise è consultabile in appendice A. Le procedure automatiche, atte a individuare il modello più parsimonioso con il miglior adattamento ai dati, portano a selezionare principalmente gli allergeni e le variabili climatiche, confermando parte dei risultati trovati dagli effetti univariati e aggiustati. Vengono invece scartati gli inquinanti, che appaiono dunque meno adatti a spiegare il fenomeno degli accessi in pronto soccorso. Si registra qualche effetto positivo e significativo anche sugli accessi per trauma, associati al valore cumulato del biossido di azoto e alla concentrazione di graminacee. Nel complesso i modelli adattati non presentano correlazione residua, dal momento che l'ACF calcolata per i residui non risulta significativa a quasi nessun ritardo (fig. 3.8).

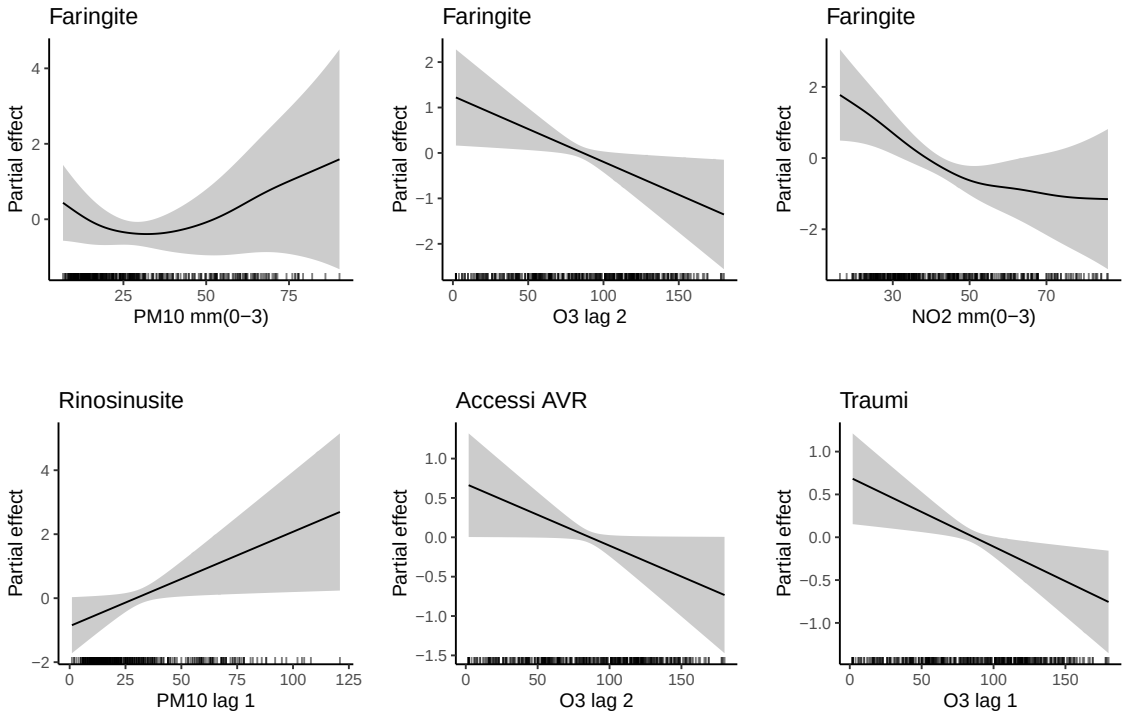


Figure 3.6: Effetti parziali significativi degli inquinanti con aggiustamento per i confondenti (dati 2020).

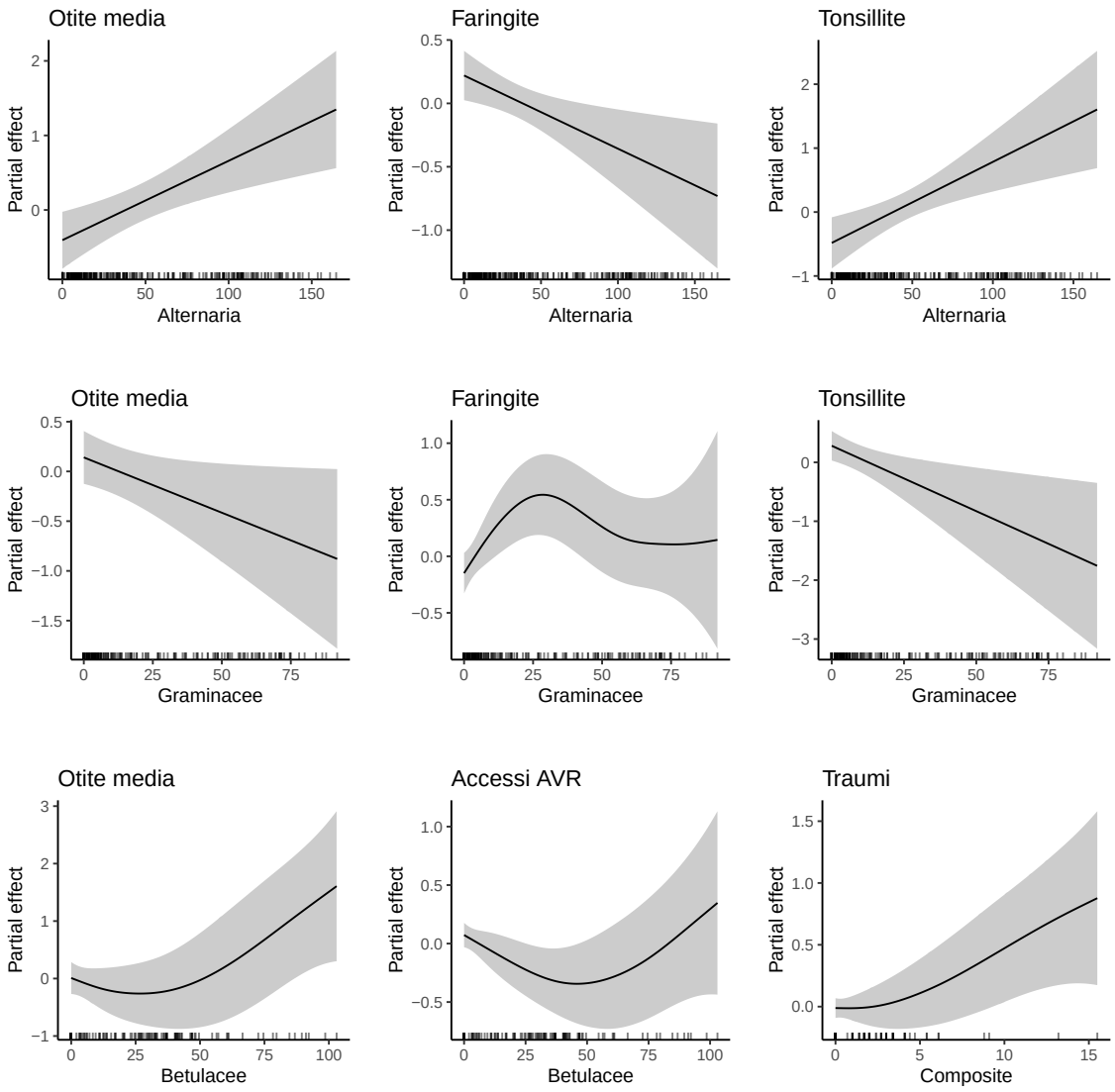


Figure 3.7: Effetti parziali significativi degli allergeni con aggiustamento per i confondenti (dati 2020).

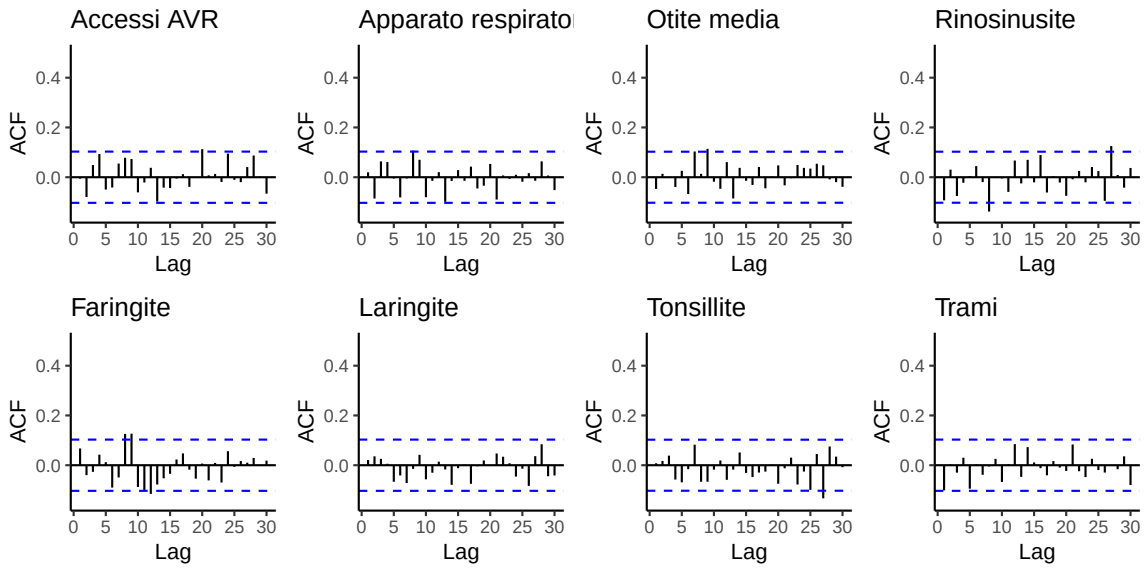


Figure 3.8: ACF residui dei modelli stepwise (dati 2020).

3.2.4 Confronto con il 2017

Rispetto a quanto emerso dalle analisi dei dati del 2020, può essere interessante verificare se l'effetto cumulato delle polveri sottili sia più intenso nel 2017, quando l'inquinamento era maggiore, tenendo conto anche del fatto che in quest'anno gli accessi in pronto soccorso erano più numerosi per l'assenza di restrizioni. A questo proposito i due dataset sono stati uniti in un formato *long*, ponendoli uno sotto l'altro in un unico dataset e definendo una nuova variabile che distinguesse i due anni. L'idea di base è verificare la presenza di interazione fra l'inquinante cumulato e l'anno e, in assenza di interazione, fornire l'effetto complessivo sui due anni.

Per tenere conto dell'eventuale presenza di effetti non lineari, viene adattata una *spline* di lisciamento separatamente per anno, tramite una formulazione del tipo:

$$\begin{aligned}\log(\mu) =& \alpha + f_1(X_M, Anno) + \beta \cdot Anno + f_2(X) + f_3(X_{-1}) + f_4(X_{-2}) + f_5(Temperatura) + \\ & f_6(Umidita\ min) + f_7(Umidita\ max) + Y_{-1}\end{aligned} \tag{3.22}$$

	Con interazione	Senza interazione
Accessi AVR	2847	2856
Apparato respiratorio	2393	2391
Otite media	1695	1702
Rinosinusite	1206	1029
Faringite	1581	1575
Laringite	1314	1310
Tonsillite	1471	1469

Table 3.5: AIC del modello GAM con e senza interazione fra valore cumulato di PM10 e anno.

La variabile risposta ritardata di un *lag* (Y_{-1}) è stata inserita per ridurre l'autocorrelazione dei residui del modello; X rappresenta il valore di polveri sottili. La rilevanza dell'interazione è stata valutata confrontando il modello completo e il modello senza interazione tramite il criterio AIC (tab. 3.5). L'AIC del modello con interazione risulta minore dell'AIC del modello senza interazione per accessi AVR, otite media e rinosinusite. Per queste diagnosi viene dunque mantenuto l'effetto distinto tra i due anni (fig. 3.9), mentre per i restanti accessi viene calcolato l'effetto complessivo (fig 3.10). Per otite media e rinosinusite si osserva che l'effetto delle polveri sottili diminuisce nel 2020 rispetto al 2017, ma non sono significativi gli effetti lineari per ciascun anno. Anche gli effetti su accessi AVR non risultano significativi, sebbene mostrino un andamento positivo in entrambi gli anni. Tra gli effetti complessivi, quelli che risultano significativi al livello 5% sono quelli su faringite e tonsillite. L'effetto sull'apparato respiratorio, rappresentato da un andamento crescente, risulta significativo solo al 10%. Conducendo un'analisi analoga per gli accessi per trauma, non si riscontra un effetto dell'inquinante cumulato né separatamente per anno, né complessivamente sui due anni. L'analisi dell'autocorrelazione dei residui risulta abbastanza soddisfacente (fig. 3.11).

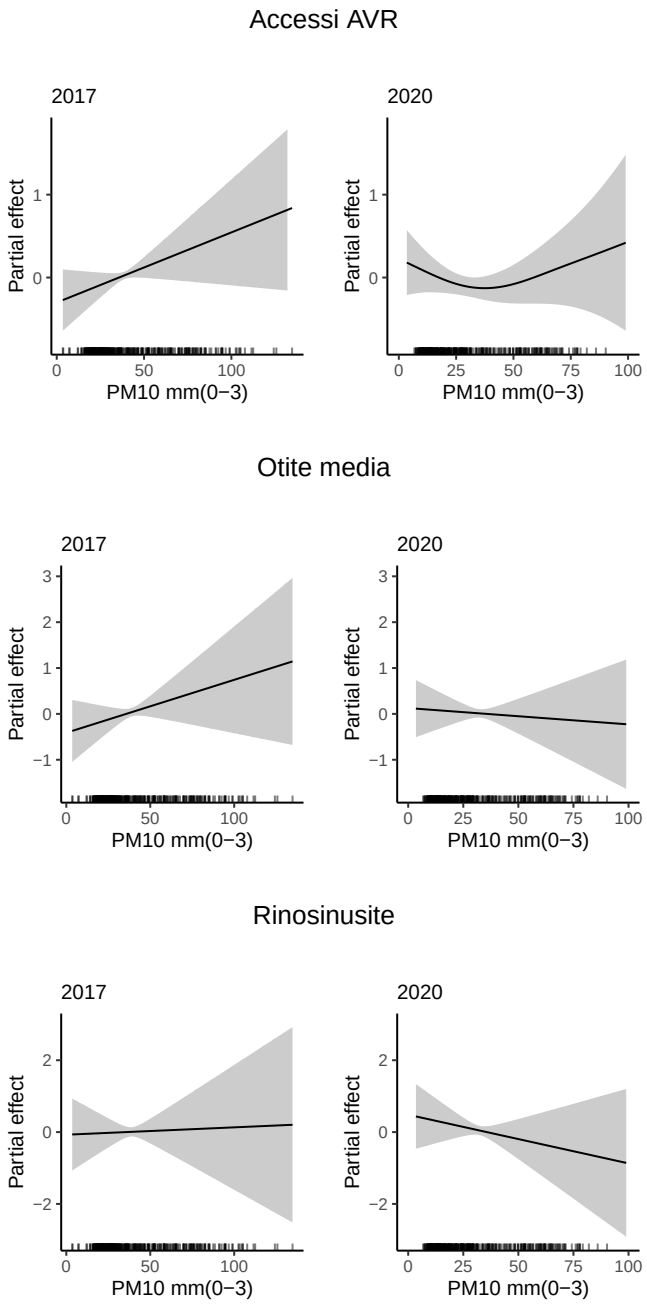


Figure 3.9: Effetto della media mobile delle polveri sottili, separatamente per anno.

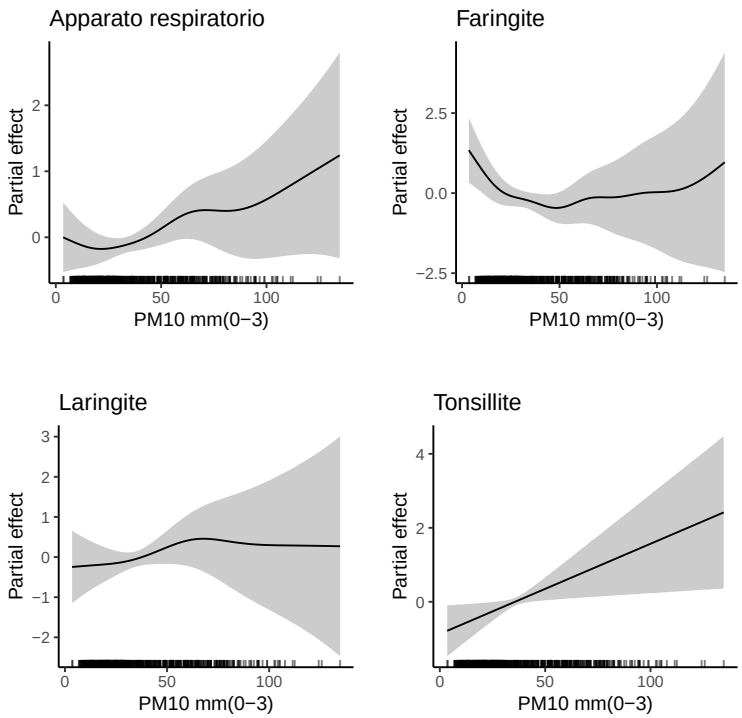


Figure 3.10: Effetto complessivo sui due anni della media mobile delle polveri sottili.

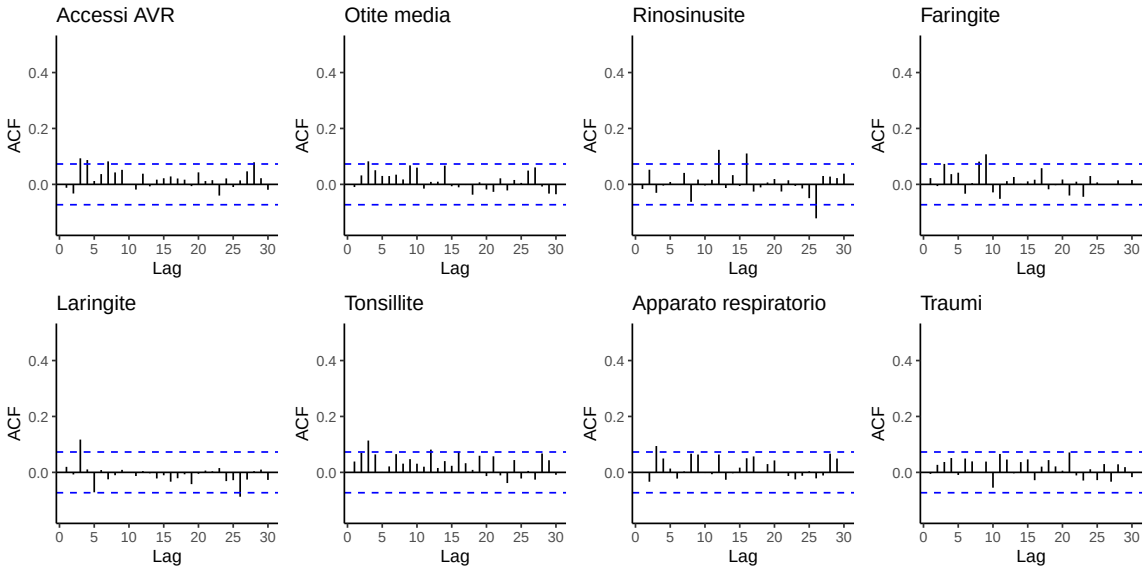


Figure 3.11: ACF dei residui dei modelli sui dati dei due anni.

3.3 Conclusioni

Nel periodo pandemico, tra gli allergeni più impattanti sulla salute si trovano l'alternaria e le betulacee. Queste ultime presentano un effetto di rischio a partire da circa 45 g/m^3, che conferma quanto indicato in letteratura per la manifestazione dei sintomi. I modelli di selezione stepwise hanno portato a selezionare principalmente caratteristiche ambientali indipendenti dalle attività dell'uomo, quali allergeni e variabili climatiche, suggerendo che queste variabili siano più adatte per spiegare il fenomeno degli accessi in pronto soccorso rispetto alla concentrazione di inquinanti. Tuttavia, analizzando separatamente l'effetto di ciascun inquinante, emerge un effetto positivo del valore medio cumulato delle polveri sottili sui 4 giorni consecutivi. L'effetto è presente in modo significativo sugli accessi per apparato respiratorio e su alcune patologie ad esso collegate, quali faringite e più marcatamente tonsillite. Anche sugli accessi totali per patologie legate alle alte vie respiratorie si osservano effetti positivi dell'inquinante cumulato, che però non risultano significativi. Ad eccezione della tonsillite, che presenta un effetto lineare, le forme funzionali dell'effetto delle polveri sottili hanno tipicamente un andamento che si conferma crescente a partire da circa 45-50

$\mu g/m^3$, in supporto alle misure di allerta adottate dall'ARPAV. Nei casi in cui l'effetto si diversifica per l'anno di analisi (ovvero per otite media, rinosinusite e accessi totali), si osserva che nel 2020 le polveri sottili hanno un impatto di minore entità rispetto al 2017 sugli accessi in pronto soccorso. È da tenere in considerazione che la ridotta numerosità del campione in analisi e la bassa frequenza di accessi, specialmente nel 2020, può aver ostacolato l'emergere in modo significativo di fattori di rischio impattanti sugli accessi.

Capitolo 4

Modelli per risposta dicotomica

4.1 Analisi statistiche

Oltre a considerare il numero di accessi giornalieri, un altro modo di studiare i dati a disposizione è analizzare la prevalenza di accessi per patologia rispetto agli accessi avvenuti per trauma, considerati come gruppo di controllo. Come operazioni preliminari, viene costruito un dataset per ogni patologia di interesse, in cui si selezionano tutti gli accessi per trauma e quelli avvenuti per la patologia in esame. L'unità statistica è il singolo accesso, pertanto oltre alle variabili ambientali (con valore identico per uno stesso giorno) si affiancano anche le variabili di sesso ed età del paziente che ha effettuato l'accesso. Viene quindi creata una variabile dicotomica che distingue gli accessi per trauma da quelli per patologia: essa costituirà la variabile risposta.

Questo tipo di analisi permette di cogliere se gli accessi per patologia si presentano in concomitanza di caratteristiche ambientali diverse rispetto agli accessi per trauma. Questo proposito risultava più difficile da perseguire nel capitolo precedente, dove concentrandosi sul numero giornaliero di accessi si studiava l'impatto dell'ambiente sull'entità delle malattie e, separatamente, sull'entità dei traumi.

4.1.1 Modelli additivi generalizzati ad effetti misti

In questo capitolo, la natura dicotomica della variabile risposta suggerisce di utilizzare come funzione di legame, per i modelli generalizzati e la loro versione additiva, la funzione *logit*

$$g(\mu) = \text{logit}(\mu) = \log\left(\frac{\mu}{1-\mu}\right) = \log\frac{P(Y=1)}{P(Y=0)} \tag{4.1}$$

dove μ è la media della variabile risposta condizionata alle esplicative.

Si può però considerare che i dati presentano una struttura gerarchica in cui la variabile di gruppo è costituita dal giorno (unità di secondo livello), mentre il singolo accesso costituisce l'unità di primo livello. La comunanza di fattori ambientali relativi allo stesso giorno può portare ad una correlazione tra gli accessi avvenuti in una stessa data, ovvero ad una concentrazione di malattia all'interno delle giornate maggiore rispetto alla casualità. Ignorare la struttura gerarchica può portare ad una valutazione errata degli *standard error* delle stime dei parametri del modello e, quindi, ad un'errata inferenza.

Per descrivere la struttura multilivello dei dati è utile l'approccio dei modelli ad effetti casuali. Sia Y_{ij} il vettore casuale della risposta con $i = 1, \ldots, n$ indice di unità di secondo livello (giorno) e $j = 1, \ldots, m_i$ indice di unità di primo livello (paziente che effettua l'accesso in pronto soccorso). Si noti che i gruppi sono di numerosità variabile (m_i), in quanto il numero di accessi può variare di giorno in giorno. Partendo dalla formulazione più generale e semplice, il modello con effetti misti per risposta normale si presenta come:

$$Y_{ij} = \mu + \alpha_i + \epsilon_{ij} \tag{4.2}$$

dove μ è l'effetto fisso e rappresenta la media generale, α_i è l'effetto dell'unità i-esima che ne sintetizza le caratteristiche, ovvero lo scostamento dell'unità i-esima dalla media generale, mentre ϵ_{ij} è un errore casuale. Si assume che $\alpha_i \sim N(0, \sigma^2_\alpha)$ e $\epsilon_{ij} \sim N(0, \sigma^2_\epsilon)$, con ϵ_{ij} indipendenti da α_i. Da questa rappresentazione si evince che le osservazioni appartenenti ad uno stesso gruppo non

sono indipendenti, in quanto

$$Cov(Y_{ij}, Y_{ih}) = E((\alpha_i + \epsilon_{ij})(\alpha_i + \epsilon_{ih})) = \sigma_\alpha^2 \tag{4.3}$$

e quindi esiste una correlazione positiva tra osservazioni dello stesso gruppo, chiamata correlazione intra-classe, data da

$$Cor(Y_{ij}, Y_{ih}) = \rho = \frac{\sigma_\alpha^2}{\sigma_\alpha^2 + \sigma_\epsilon^2} \tag{4.4}$$

e rappresenta la quota di varianza totale attribuibile alle differenze fra gruppi.

Quello appena presentato è il modello lineare normale con intercetta casuale; aggiungendo al modello anche eventuali covariate, si ottiene il modello lineare normale ad effetti misti:

$$Y_{ij} = x_{ij}\beta + z_{ij}u_i + \epsilon_{ij} \tag{4.5}$$

in cui p variabili esplicative x_{ij} sono associate ad effetti fissi (β vettore p-dimensionale), e q variabili esplicative z_{ij} sono associate ad effetti casuali (u_i vettore q-dimensionale con distribuzione $N_q(0, \Sigma_u)$). Gli effetti fissi sono identificati da parametri relativi all'intera popolazione, mentre gli effetti casuali sono identificati da parametri gruppo-specifici. Si assume l'indipendenza tra errori casuali ϵ_{ij} ed effetti casuali u_i, nonché l'indipendenza tra effetti casuali relativi a diverse unità di secondo livello (u_i indipendente da u_k per $i \neq k$) e tra errori casuali relativi alla stessa unità e a diverse unità. In un modello a effetti misti, le intercette casuali catturano la variazione sulla risposta dovuta ai gruppi, mentre le pendenze casuali catturano la variazione, dovuta alla presenza di gruppi, dell'effetto dei predittori sulla risposta, e consentono quindi differenze nella relazione tra il predittore e la variabile risposta per diversi gruppi.

Per estendere tale modello ad un tipo di risposta non normale si assume che, condizionatamente agli effetti casuali, le osservazioni sulla risposta Y_{ij} siano distribuite secondo un modello lineare generalizzato con effetti misti (*Generalized Linear Mixed Model*, GLMM):

$$g(E(Y_{ij}|u_i)) = x_{ij}\beta + z_{ij}u_i \tag{4.6}$$

dove, nel caso di risposta binaria, la funzione di legame $g(\cdot)$ è la funzione $logit(\cdot)$. Il modello così formulato viene anche chiamato modello logistico-normale con effetti misti. In questo caso una misura della correlazione intraclasse, interpretabile come percentuale di variabilità dovuta alla struttura gerarchica, si può calcolare come segue:

$$\rho = \frac{\sigma_\alpha^2}{\sigma_\alpha^2 + \pi^2/3} \tag{4.7}$$

La versione additiva del GLMM è un modello additivo generalizzato ad effetti misti (*Generalized Additive Mixed Models*, GAMM)

$$g(E(Y|u)) = \sum_{j=1}^{p} f_j(x_j) + \sum_{k=1}^{q} z_k u_k \tag{4.8}$$

che può essere altresì visto come una versione ad effetti misti del modello GAM. Oltre a permettere la presenza di intercette e pendenze casuali, il modello GAMM permette di cogliere effetti non lineari dovuti alla presenza dei gruppi. Per la stima del modello, gli effetti casuali vengono considerati come lisciatori e subiscono uno *shrinkage* come parte del processo di massimizzazione della verosimiglianza penalizzata. Per rendere l'idea di come un modello GAMM può adattarsi ai dati, si presenta in fig. 4.1 un esempio su dati simulati (codice consultabile in appendice).

4.1.2 Foresta casuale ad effetti misti

Finora sono stati utilizzati metodi che permettono di avere un'interpretabilità dei risultati, in modo da determinare i fattori di rischio sulla salute pubblica. Può essere però di interesse esplorare anche metodi meno interpretabili ma che possono essere di sostegno ai risultati emersi. Inoltre, può essere interessante anche considerare le interazioni tra i fattori ambientali studiati. La foresta casuale (*Random Forest*) è una combinazione di alberi (*Classification And Regression Tree*, CART) che fanno uso di diversi sottoinsiemi casuali delle covariate disponibili. La foresta casuale non fornisce una direzione né una misura dell'associazione fra le variabili esplicative e la risposta, ma dà una misura dell'importanza di ciascuna esplicativa nel predire la risposta. Questo metodo può

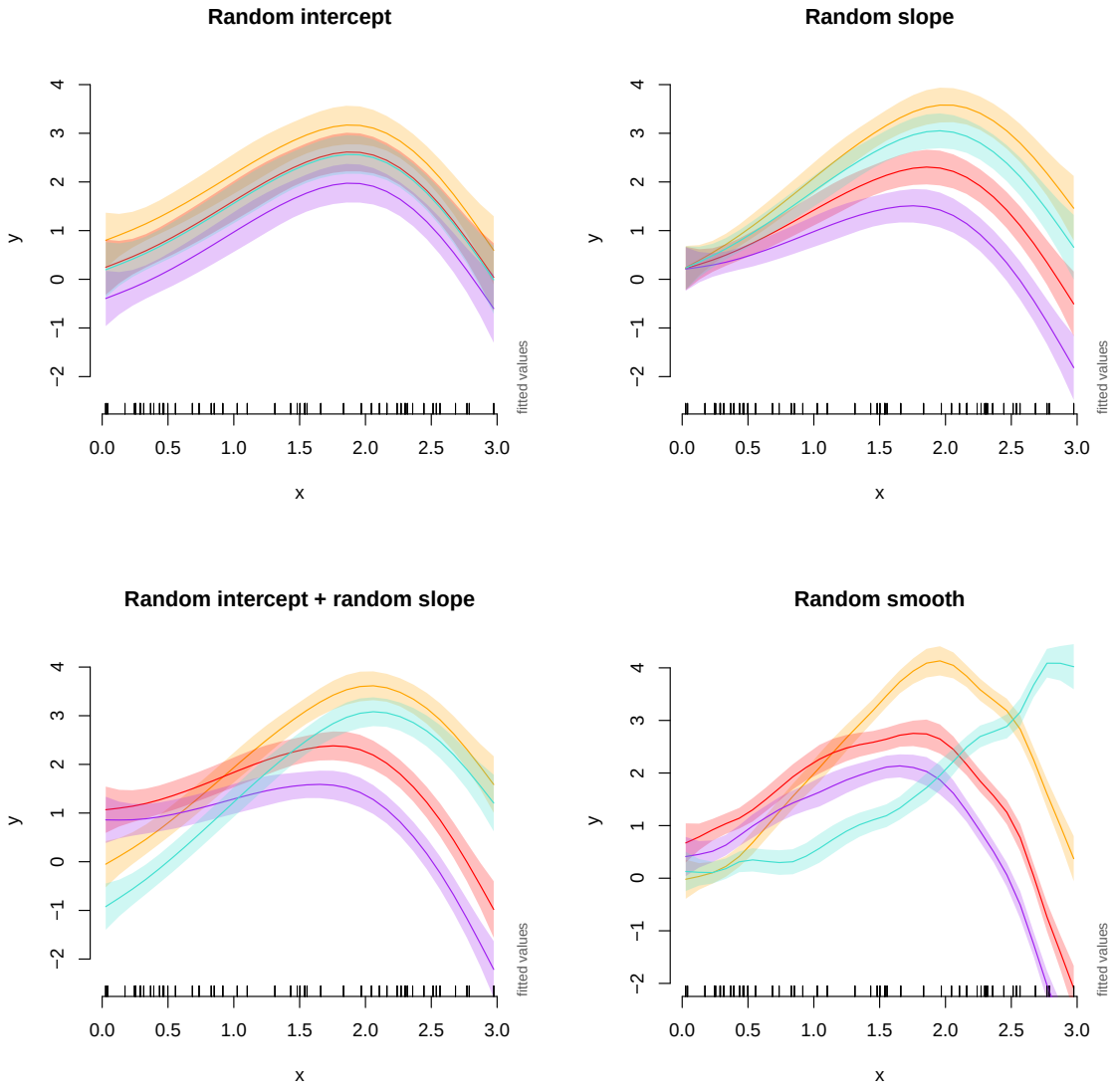


Figure 4.1: Esempio su dati simulati del modello GAMM.

essere utilizzato per la classificazione di risposta binaria e può essere inoltre esteso a dati con struttura gerarchica, secondo quanto proposto ad esempio da Pellagatti et al. [2021]. Riprendiamo la formulazione di un modello ad effetti misti in cui il vincolo lineare viene sostituito da una funzione più flessibile (f):

$$\begin{aligned} \mu_i &= E(y_i|u_i) \\ g(\mu_i) &= \eta_i = f(x_i) + z_i u_i \\ u_i &\sim N_q(0, \Sigma_u) \end{aligned} \tag{4.9}$$

L'idea alla base dell'algoritmo GMERF (*Generalized Mixed-Effects Random Forest*) è di stimare gli effetti fissi $f(x_i)$ con una foresta casuale e gli effetti casuali $z_i u_i$ con un modello ad effetti misti. In particolare si osserva che se gli effetti casuali fossero noti, allora gli effetti fissi così definiti

$$f(x_i) = \eta_i - z_i u_i \tag{4.10}$$

possono essere considerati come una variabile dipendente su cui adattare una foresta casuale. Similmente, se gli effetti fissi $f(x_i)$ fossero noti, allora gli effetti casuali così definiti

$$z_i u_i = \eta_i - f(x_i) \tag{4.11}$$

possono essere stimati con un modello ad effetti misti. Le due fasi di stima vengono quindi alternate nell'algoritmo fino a convergenza. Quest'ultima si può basare sulla differenza della log-verosimiglianza del modello a effetti misti tra due iterazioni successive, oppure sulla differenza tra gli effetti casuali stimati tra due iterazioni successive. Per la stima di $f(x_i)$ nella prima iterazione, il valore iniziale di η_i viene stimato da un GLM standard utilizzando tutte le covariate come effetti fissi, e gli effetti casuali vengono considerati nulli. In appendice è esposto il codice per l'implementazione dell'algoritmo GMERF.

Patologia	Numerosità tot.	Accessi patologia	N. giorni	N. giorni patologia
Rinosinusite	815	84 (10%)	324	71
Otite media	889	158 (18%)	328	117
Faringite	1003	272 (27%)	335	176
Laringite	822	91 (11%)	330	75
Tonsillite	843	112 (13%)	330	89
Apparato resp.	1215	484 (40%)	349	261
Accessi AVR	1457	726 (50%)	355	303

Table 4.1: Informazioni sul dataset creato per ciascuna patologia.

4.2 Risultati

4.2.1 Analisi esplorative

Per le frequenze assolute di accessi per ciascuna patologia e per i traumi si rimanda alla tabella 2.1 del capitolo 2. Gli accessi per trauma sono in totale 731, e vengono selezionati in ciascun dataset come gruppo di controllo. In tabella 4.1, a fini riassuntivi, viene riportata la numerosità totale del dataset creato per ciascuna patologia e la percentuale di accessi dovuti a quest'ultima. Viene riportato anche il numero di giorni coinvolti (non in tutte le giornate infatti si registra un accesso per trauma o per la patologia in esame), e il numero di giorni in cui è avvenuto almeno un accesso per la patologia. Poiché il campione di controllo è molto più numeroso rispetto agli altri tipi di accessi, la variabile risposta risulta sbilanciata in tutti i dataset (ad eccezione di quello completo). Inoltre, gli accessi per causa specifica si presentano in un numero relativamente ridotto di giornate, soprattutto quelli per rinosinusite, laringite e tonsillite.

In tabella 4.2 viene riportata per ciascuna patologia la percentuale di donne e l'età mediana (primo e terzo quartile) dei soggetti che hanno effettuato l'accesso per quella causa. La distribuzione dell'età distintamente per diagnosi è rappresentata graficamente in figura 4.2. Si osserva che l'età mediana si aggira intorno ai 40 anni per otite, rinosinusite e faringite; la tonsillite colpisce soggetti più giovani, mentre la laringite soggetti leggermente più anziani. Uomini e donne sembrano abbas-

	% Femmine	Età mediana (IQR)
Otite media	65 (41%)	40 (21, 55)
Rinosinusite	35 (42%)	42 (32, 58)
Faringite	144 (53%)	40 (26, 54)
Laringite	45 (49%)	51 (36, 68)
Tonsillite	56 (50%)	24 (16, 30)
Traumi	343 (47%)	47 (16, 71)

Table 4.2: Distribuzione di sesso ed età stratificata per diagnosi.

tanza equidistribuiti; una prevalenza più marcata di soggetti maschi si osserva per otite e rinosinusite.

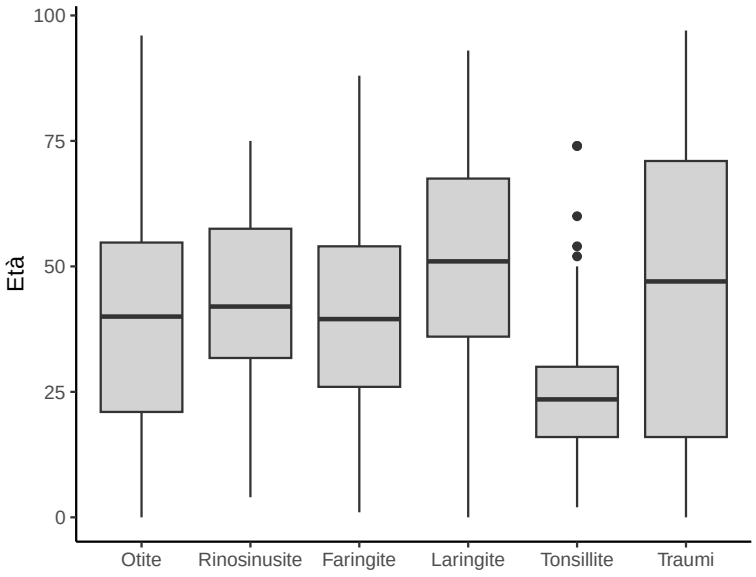


Figure 4.2: Distribuzione di età stratificata per diagnosi.

4.2.2 Effetti univariati

Analogamente a quanto svolto per l'analisi su dati di conteggio (cap. 3), vengono presentati innanzitutto i risultati sugli effetti univariati. I risultati dei modelli logistici, espressi in termini di rapporto fra quote (*odds ratio*), sono esposti in tab. 4.3, e sono associati ad un incremento di 10 unità della variabile esplicativa (nel caso di esplicative continue). A titolo esemplificativo, si interpreta l'effetto dell'alternaria sugli accessi AVR: l'incremento di 10 spore/m^3 della concentrazione di altenaria comporta un aumento del 5% dell'*odd* della probabilità di accedere per patologie legate alle alte vie respiratorie rispetto ad accedere per trauma.

La tabella 4.4 riporta i *p-value* associati a ciascun effetto non parametrico per le esplicative quantitative ottenuto dai modelli GAM univariati con funzione di legame *logit*. In figura 4.3 vengono rappresentati i grafici degli effetti significativi ($p\text{-}value < 0.05$).

4.2.3 Effetti ritardati e aggiustamento

Analogamente a quanto svolto per l'analisi per dati di conteggio, l'analisi per risposta dicotomica si è focalizzata sui modelli additivi. Sono stati quindi adattati modelli GAM su ciascuna variabile risposta e separatamente per ogni inquinante, considerando anche gli effetti ritardati e l'aggiustamento per le variabili meteorologiche e i possibili confondenti, oltre che le variabili *baseline* del singolo paziente (sesso ed età). I modelli GAM sono stati poi utilizzati per studiare l'effetto degli allergeni con l'aggiustamento per le variabili esplicative *baseline*, meteorologiche e confondenti. I confondenti qui considerati sono il giorno della settimana (da lunedì a domenica, variabile quantitativa discreta), il giorno festivo (dicotomica), l'impatto delle restrizioni (quantitativa continua).

I modelli per lo studio dell'effetto degli inquinanti si formulano come:

$$\begin{aligned}\text{logit}(\mu) =& \alpha + f_1(X) + f_2(X_{-1}) + f_3(X_{-2}) + f_4(X_M) + f_5(Temperatura) + f_6(Umidita\ min) + \\ & f_7(Umidita\ max) + \beta_1 Wday + \beta_2 Festivo + f_8(Restrizioni) + f_9(Eta) + \beta_3 Sesso\end{aligned} \tag{4.12}$$

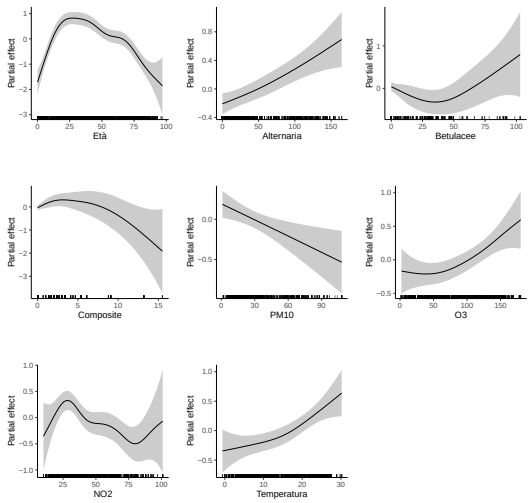
dove X è l'inquinante. In figura 4.4 vengono riportati gli effetti significativi riscontrati degli in-

	AVR	App. resp.	Otite media	Rinosinusite	Faringite	Laringite	Tonsillite
Età 26-50	**2.139** ***	**2.151** ***	**1.716** *	**3.351** ***	**2.899** ***	**4.085** ***	1.029
Età 51-75	1.142	0.953	1.189	**2.665** **	1.252	**3.337** ***	**0.101** ***
Età 76+	**0.359** ***	**0.460** ***	**0.190** ***	NS	**0.511** *	2.110	NS
Maschio	0.966	0.848	1.265	1.238	0.786	0.904	0.884
Alternaria	**1.052** ***	**1.049** ***	**1.067** ***	1.047	1.030	**1.053** *	**1.099** ***
Betulacee	0.988	0.950	1.078	0.966	0.917	**1.125** *	**0.802** *
Composite	0.918	0.845	0.782	1.673	0.756	0.251	1.731
Coriacee	0.967	0.957	1.001	0.950	**0.927** *	1.075	0.898
Graminacee	0.986	0.990	0.969	0.992	1.037	1.020	**0.797** **
PM10	**0.935** **	**0.936** *	0.948	0.915	**0.919** *	0.986	0.948
O_3	**1.044** ***	**1.045** ***	**1.045** *	1.042	**1.058** ***	1.052	1.012
NO_2	**0.904** ***	0.905 ***	**0.882** **	0.948	**0.878** ***	**0.862** *	1.004
Temperatura	**1.031** ***	**1.033** ***	**1.029** *	1.030	**1.041** ***	1.027	1.023
Umidità min.	0.996	0.996	0.998	0.991	0.998	0.993	0.994
Umidità max.	**0.982** *	**0.982** *	0.989	0.982	1.001	**0.956** ***	**0.969** *
Precipitazioni	0.992	0.977	1.019	0.981	0.989	0.969	0.950

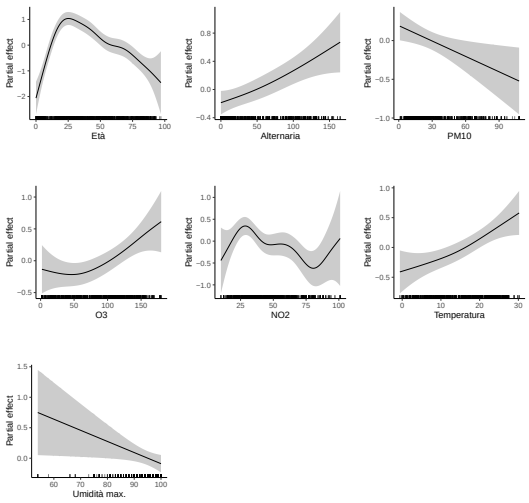
Table 4.3: *Odds ratio* ottenuti dai modelli logistici univariati per la regressione della probabilità di accesso sulle variabili ambientali. Risultati associati a un incremento di 10 unità per le variabili esplicative continue. Le modalità di riferimento per sesso ed età sono rispettivamente 'Femmina' e '0-25'. Gli asterischi indicano la significatività: *** per $p.value < 0.001$, ** per $0.001 \leq p.value < 0.01$, * per $0.01 \leq p.value < 0.05$. NS = Non Stimabile.

	Accessi AVR	Otite media	Rinos.	Faring.	Laring.	Tons.	Appar. resp.
Età	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**0.001**	**<0.001**	**<0.001**
Alternaria	**<0.001**	**0.002**	0.065	0.061	0.057	**<0.001**	**0.001**
Betulacee	**0.043**	**0.001**	0.756	0.078	**0.024**	**0.034**	0.244
Composite	**0.033**	0.456	**0.021**	0.236	0.542	0.194	0.165
Coriacee	0.221	0.343	0.362	0.054	0.156	**0.041**	0.105
Graminacee	0.065	0.446	0.439	**0.006**	0.687	**0.002**	0.071
PM10	**0.005**	0.286	0.125	**0.011**	0.775	0.380	**0.013**
O_3	**0.001**	**0.048**	0.129	**0.002**	**0.043**	0.622	**0.003**
NO_2	**0.001**	**0.026**	0.376	**0.003**	**0.015**	0.946	**0.002**
Temperatura	**<0.001**	**0.010**	0.053	**<0.001**	0.127	0.124	**<0.001**
Umidità min.	0.211	0.793	0.287	0.860	0.396	0.329	0.355
Umidità max.	0.059	**0.014**	0.180	0.632	**<0.001**	**0.019**	**0.034**
Precipitazioni	0.485	0.162	0.498	0.379	0.638	0.128	0.489

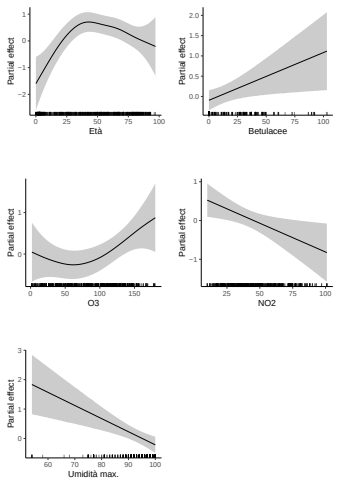
Table 4.4: *p-value* degli effetti ottenuti dai modelli GAM univariati.



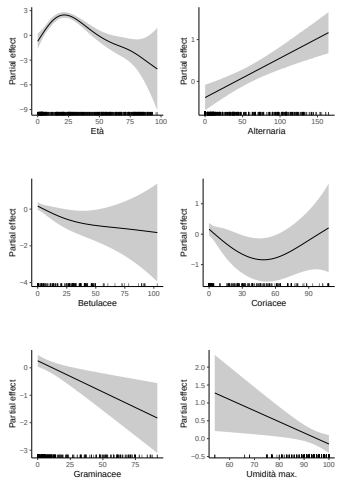
(a) Accessi AVR



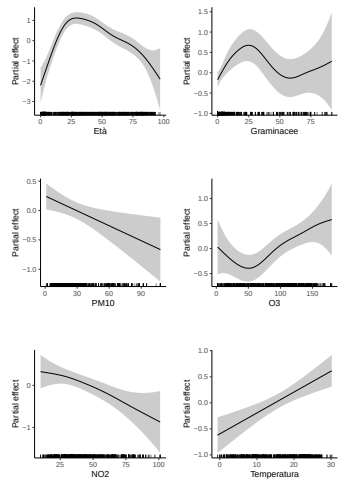
(b) Apparato respiratorio



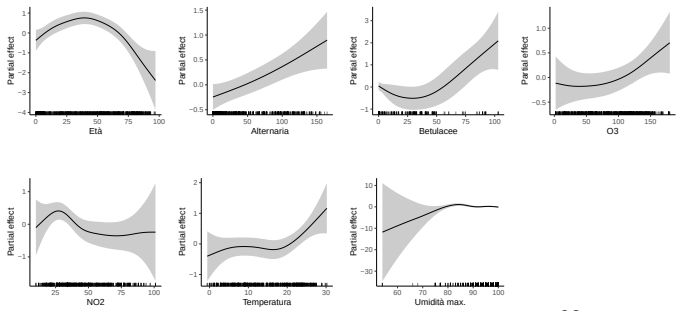
(c) Laringite



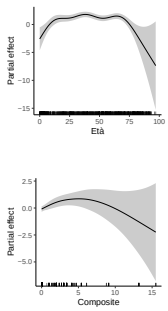
(d) Tonsillite



(e) Faringite



(f) Otite media



(g) Rinosinus.

Figure 4.3: Effetti significativi ottenuti dai modelli GAM univariati per ciascuna diagnosi.

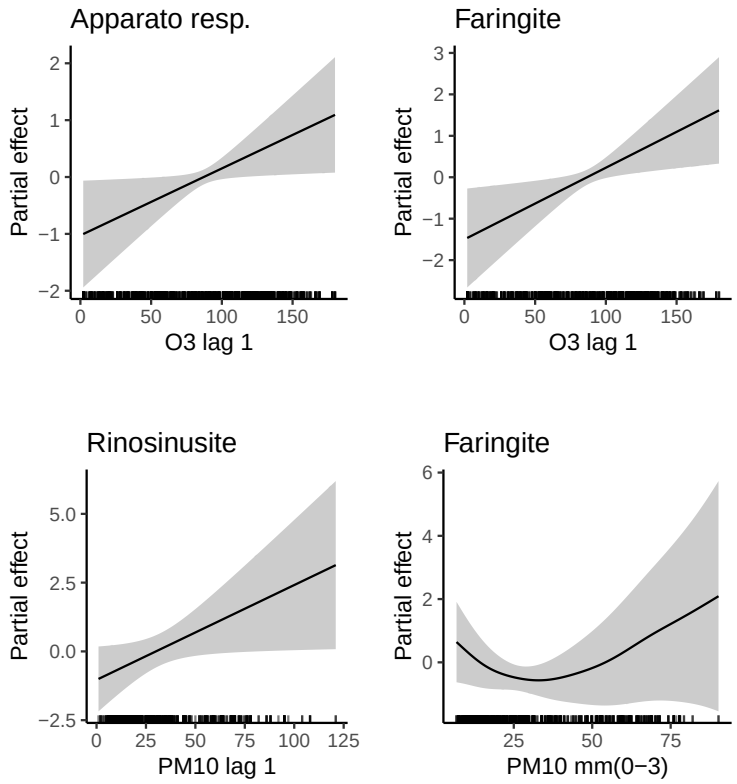


Figure 4.4: Effetti parziali significativi degli inquinanti con aggiustamento per i confondenti (dati 2020).

quinanti dopo l'aggiustamento per le covariate sopra specificate.

I modelli per lo studio dell'effetto degli allergeni si formulano come:

$$\begin{aligned}\text{logit}(\mu) =& \alpha + f_1(X) + f_2(Temperatura) + f_3(Umidita\ min) + f_4(Umidita\ max) + \\ & \beta_1 Wday + \beta_2 Festivo + f_5(Restrizioni) + f_6(Eta) + \beta_3 Sesso\end{aligned} \tag{4.13}$$

dove X è l'allergene. In figura 4.5 vengono riportati gli effetti significativi riscontrati degli allergeni dopo l'aggiustamento.

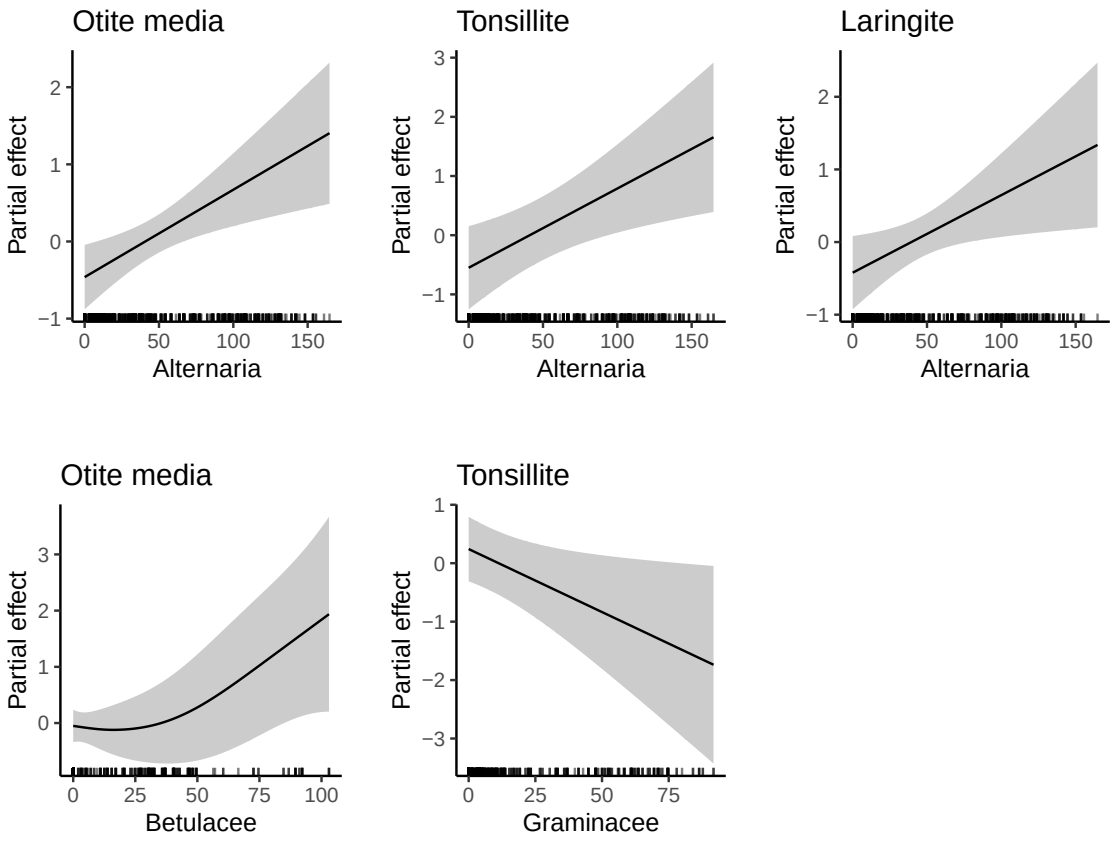


Figure 4.5: Effetti parziali significativi degli allergeni con aggiustamento per i confondenti (dati 2020).

Diagnosi	ICC
Otite media	0.152
Rinosinusite	0.153
Faringite	0.066
Laringite	0.365
Tonsillite	0.165
Apparato respiratorio	0.042
Accessi AVR	0.045

Table 4.5: Correlazione intraclasse ottenuta dai modelli con solo intercetta casuale.

4.2.4 Giorno dell'anno come variabile di gruppo

In tabella 4.5 viene riportata per ciascuna variabile risposta la correlazione intraclasse ottenuta dal modello con sola intercetta casuale per il giorno dell'anno. La percentuale di variabilità dovuta alla presenza di gruppi appare rilevante soprattutto per le diagnosi di laringite, tonsillite, otite e rinosinusite, mentre è prossima a 0 per faringite e le macrodiagnosi. Adattando modelli GAMM ad intercetta casuale per la giornata, si confermano i risultati significativi emersi dai modelli GAM: tenere conto della struttura gerarchica non sembra dunque apportare differenze alle conclusioni inferenziali.

Il modello *random forest* con intercetta casuale è stato adattato separatamente per ciascuna diagnosi. Le covariate considerate comprendono anche gli effetti ritardati degli inquinanti. Sono state implementate foreste di 300 alberi, con una selezione casuale di 5 variabili esplicative per ciascun albero. Il numero massimo di iterazioni è stato fissato a 30 e la tolleranza per definire la convergenza (differenza tra la log-verosimiglianza fra due iterazioni successive) è stata impostata inizialmente a 0.01. Poiché l'algoritmo non sembrava convergere, è stato alzato il numero di iterazioni. Nei grafici in figura 4.6 viene mostrato a titolo esemplificativo il valore della log-verosimiglianza per ciascuna iterazione dell'algoritmo GMERF sulla risposta che distingue gli accessi per trauma da quelli per tonsillite. Dal grafico con 100 iterazioni, si osserva chiaramente che a partire da circa 20 iterazioni il trend si stabilizza, segno che la convergenza viene raggiunta, e quindi il mancato ar-

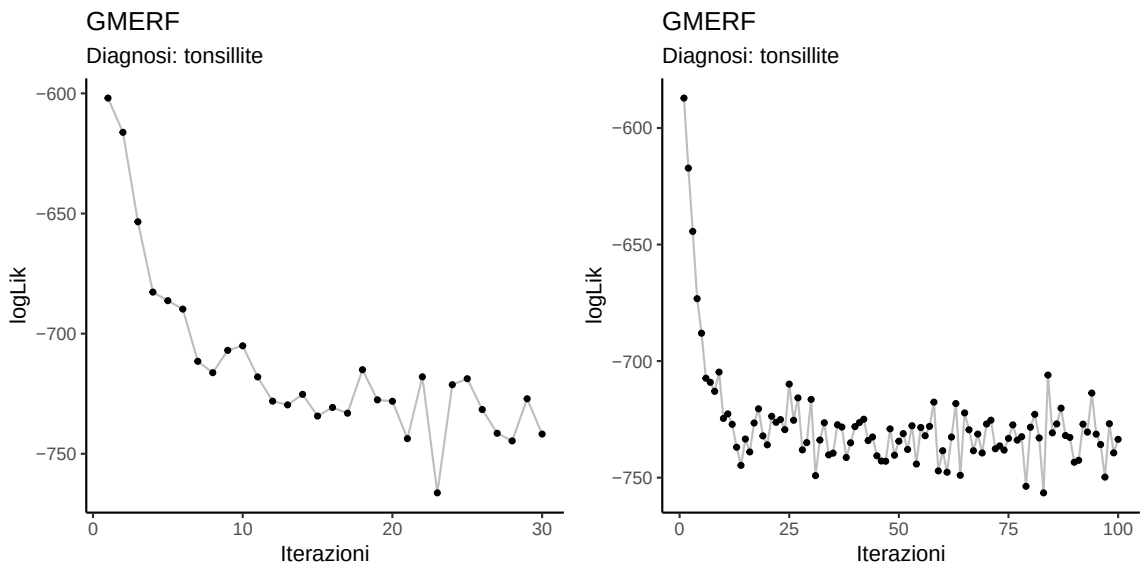


Figure 4.6: Log-verosimiglianza del modello a effetti misti ottenuta per ogni iterazione dell'algoritmo GMERF.

resto dell'algoritmo era dovuto a una soglia di tolleranza troppo bassa: quest'ultima è stata quindi aumentata a 1. In tabella 4.6 viene riportato il numero di iterazioni necessarie per raggiungere la convergenza, da cui è interessante notare come il numero di iterazioni sia minore per le diagnosi che presentavano una maggiore correlazione intraclasse. In figura 4.7 vengono mostrati per ciascuna diagnosi i grafici sull'importanza delle variabili (basata sull'incremento dell'errore quadratico medio quando i valori della variabile corrispondente vengono permutati); per semplicità vengono mostrate solo le prime 10 variabili esplicative più importanti. Le variabili soggetto-specifiche (sesso ed età) sono quelle più rilevanti per predire la risposta; seguono tipicamente la temperatura e l'alternaria, anche se generalmente le variabili ambientali riportano misure di importanza molto simili tra loro.

4.2.5 Confronto con il 2017

Un primo obiettivo è quello di studiare come cambia l'effetto del valore cumulato delle polveri sottili nei due anni. Secondariamente, in base a quanto emerso dai risultati precedenti, può essere interessante valutare anche come cambia l'effetto dell'ozono ritardato di un giorno. Per svolgere

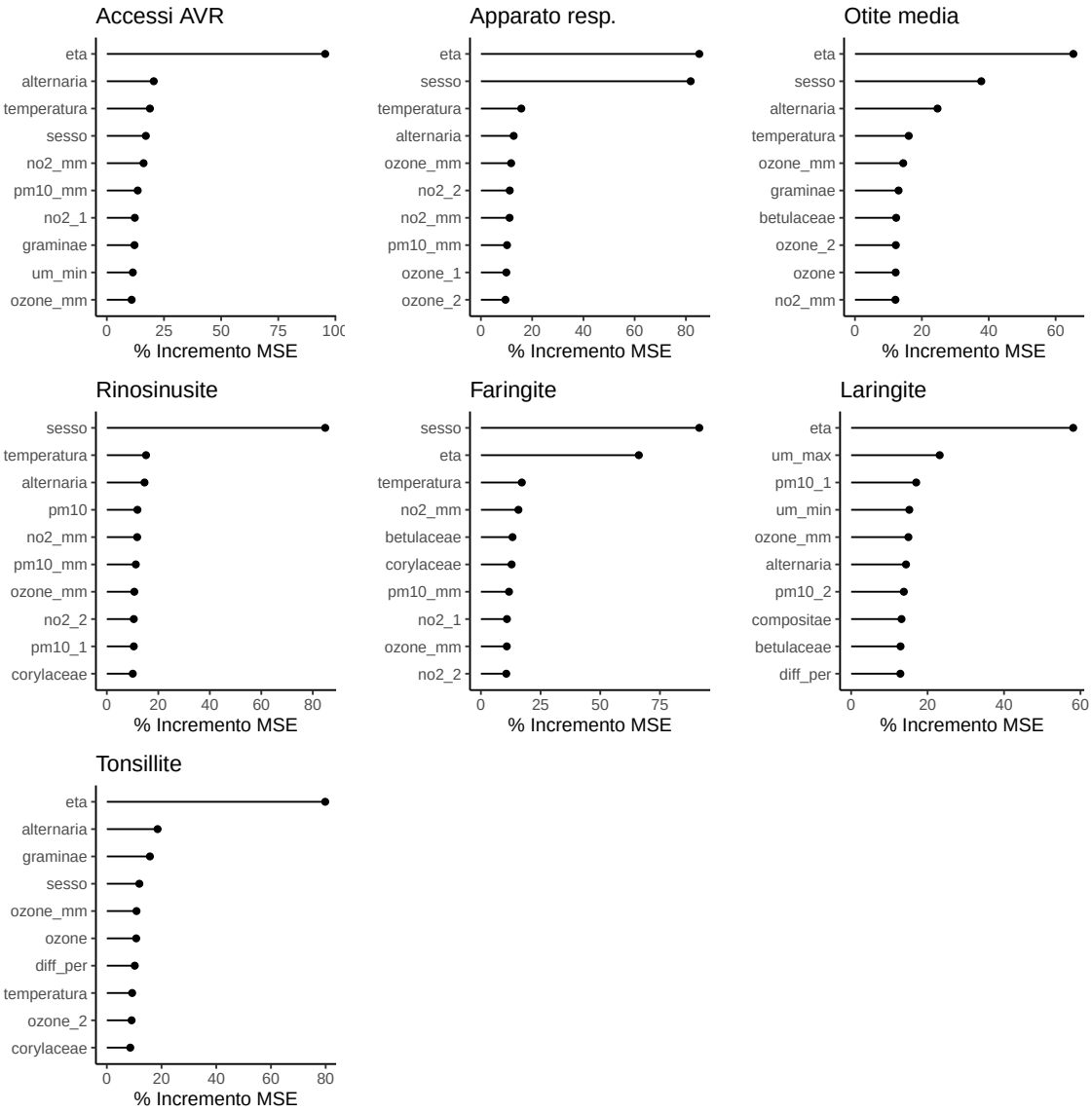


Figure 4.7: Importanza delle variabili dal modello GMERF.

Diagnosi	Numero iterazioni
Otite media	13
Rinosinusite	33
Faringite	22
Laringite	4
Tonsillite	9
Apparato respiratorio	50
Accessi AVR	25

Table 4.6: Numero di iterazioni dell'algoritmo GMERF necessarie per raggiungere la convergenza.

Diagnosi	ICC 2020	ICC 2017	ICC 2020+2017
Otite media	0.152	0.041	0.095
Rinosinusite	0.153	n.s.	0.046
Faringite	0.066	0.121	0.102
Laringite	0.365	0.045	0.132
Tonsillite	0.165	0.008	0.067
Apparato respiratorio	0.042	0.003	0.023
Accessi AVR	0.045	0.0003	0.030

Table 4.7: Correlazione intraclasse ottenuta dai modelli con solo intercetta casuale. n.s. = "non stimabile".

il confronto, è stato definito un dataset per ciascuna diagnosi anche con i dati del 2017 in modo analogo a quanto fatto per il 2020 (ovvero considerando tutti gli accessi per diagnosi e per trauma); i due dataset riferiti a ciascuna diagnosi sono stati poi collocati uno sotto l'altro, definendo una variabile dicotomica che distinguesse i due anni. La tabella 4.7 riprende le correlazioni intraclasse calcolate sui dati del 2020 (tab. 4.5) e vi affianca le correlazioni intraclasse sui dati del 2017 e sui dati che considerano i due anni insieme. Le correlazioni intraclasse nel 2017 risultano tipicamente prossime a 0.

Se si tiene della struttura gerarchica nei dataset con entrambi gli anni, si hanno circa il doppio

	Con interazione	Senza interazione
Accessi AVR	5170	6146
Apparato respiratorio	4016	4739
Otite media	2499	2693
Rinosinusite	1687	1747
Faringite	2219	2349
Laringite	1844	1922
Tonsillite	2143	2274

Table 4.8: AIC del modello GAM con e senza interazione fra valore cumulato di PM10 e anno.

del numero di giornate e, quindi, del numero di gruppi per cui stimare un'intercetta casuale. Si è visto che nei modelli GAMM questo comporta un elevato costo computazionale, inoltre dalle analisi precedenti tenere conto della struttura gerarchica nei modelli GAM non andava a cambiare le conclusioni inferenziali: sono stati pertanto utilizzati modelli GAM per studiare l'interazione fra l'anno e il valore cumulato delle polveri sottili inserendo una *spline* per tale effetto separatamente per anno. Nei modelli sono stati inseriti come variabili di aggiustamento le variabili climatiche e i valori ritardati delle polveri sottili. Confrontando gli AIC ottenuti dai modelli con e senza interazione, si osserva che i modelli con interazione presentano un adattamento migliore (tab 4.8). Gli effetti riscontrati non risultano significativi (ad eccezione dell'effetto sulla faringite nel 2020), anche se a livello grafico si possono vedere delle differenze nei due anni (fig. 4.8). In modo analogo viene condotta un'analisi per l'ozono. In tabella 4.9 vengono riportati gli AIC dei modelli con e senza interazione con l'anno, portando a preferire i modelli che adattano una *spline* separatamente per anno. Per tutte le diagnosi si osserva che l'effetto dell'ozono nel 2020 è più marcato rispetto a quanto accade nel 2017 (fig. 4.9). Ad eccezione delle diagnosi di laringite e tonsillite, inoltre, gli effetti nel 2020 risultano significativi, al contrario di quanto accade nel 2017.

La struttura gerarchica è invece stata considerata con il modello GMERF. Con l'obiettivo di valutare se l'effetto del valore cumulato delle polveri sottili sia diverso nei due anni di studio, è stata inserita l'interazione fra anno e valore cumulato delle polveri sottili (in aggiunta agli effetti

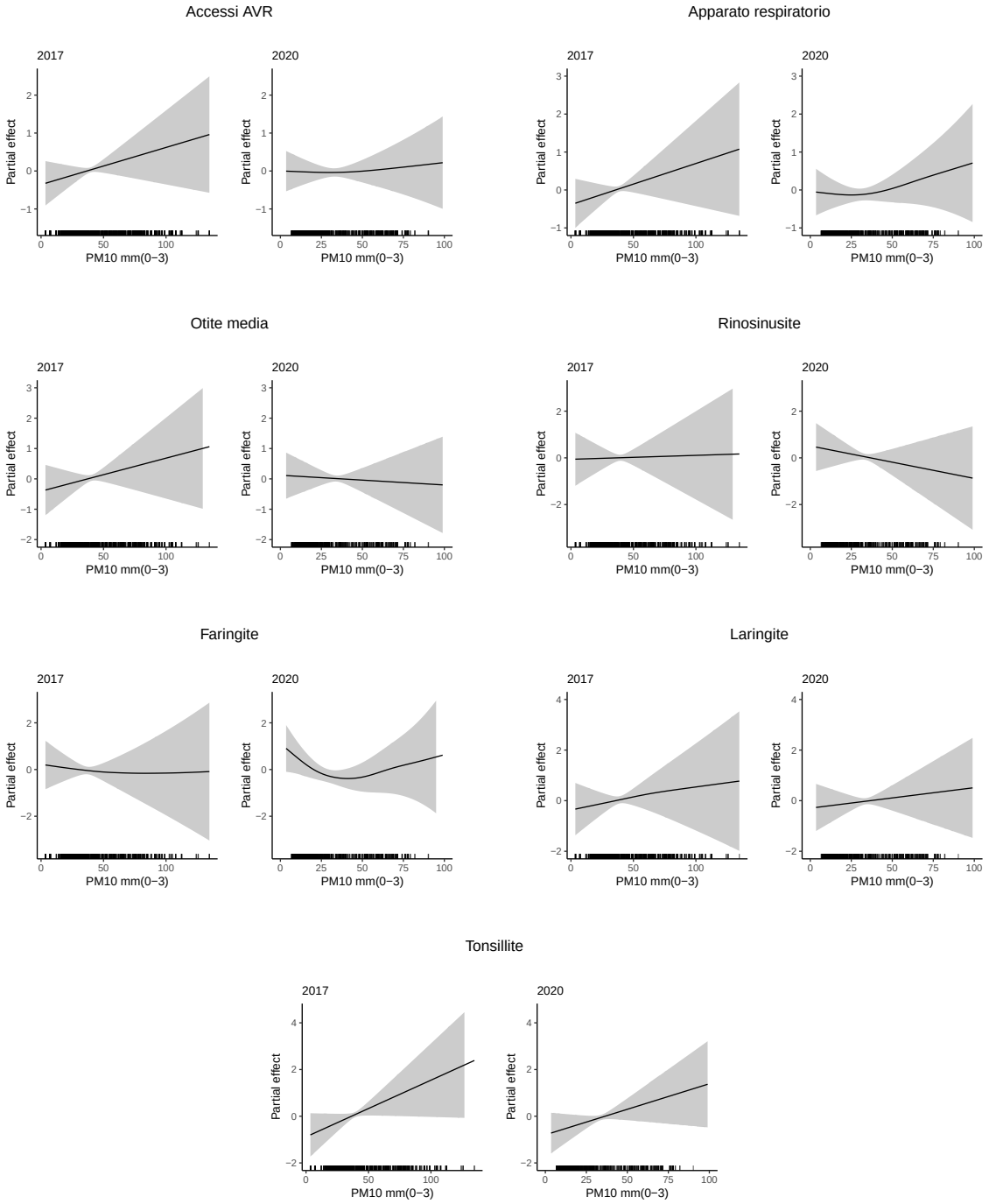


Figure 4.8: Effetti della media mobile delle polveri sottili.

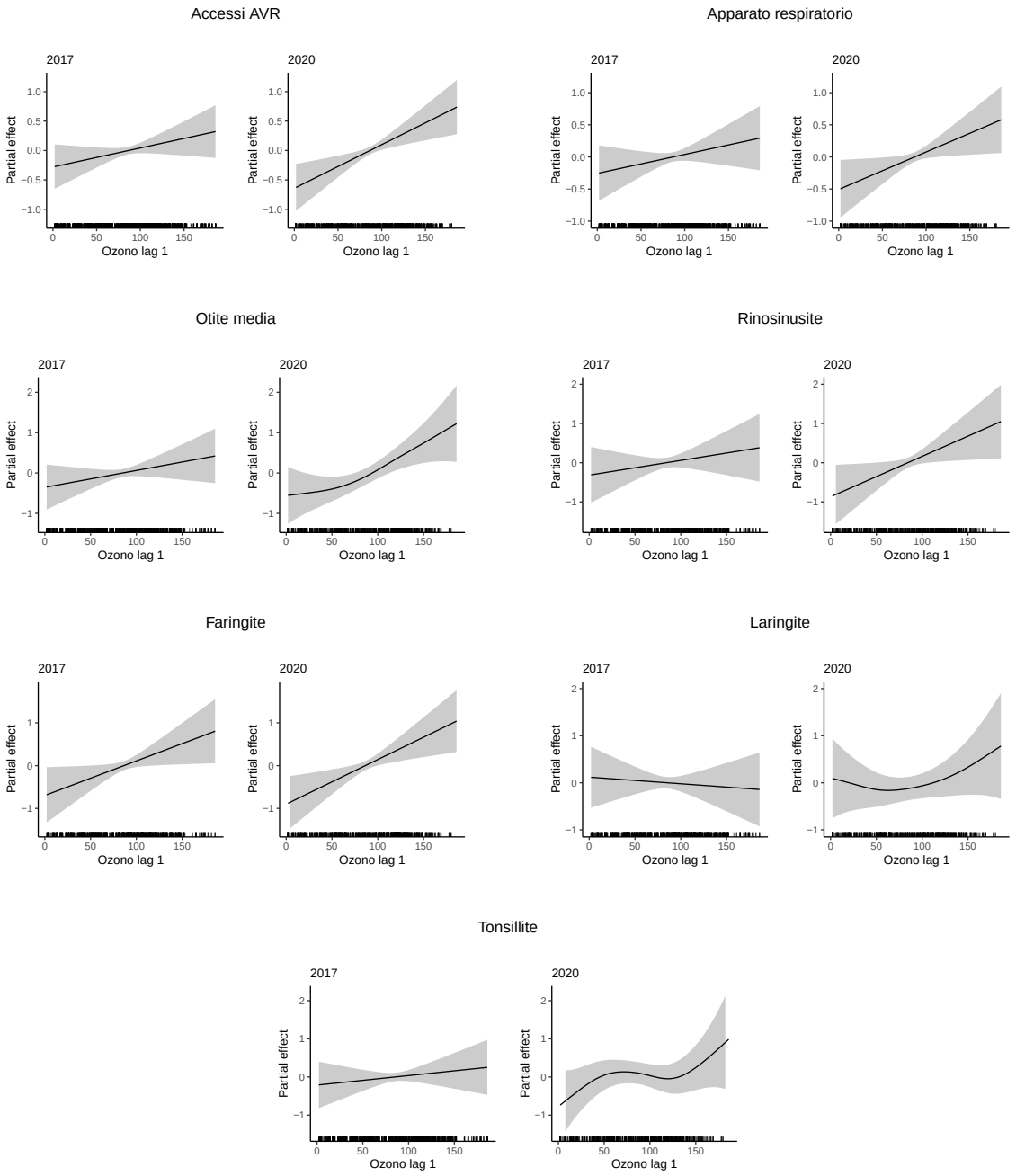


Figure 4.9: Effetti dell'ozono ritardato di un giorno.

	Con interazione	Senza interazione
Accessi AVR	5165	6844
Apparato respiratorio	4014	4738
Otite media	2501	2694
Rinosinusite	1682	1744
Faringite	2219	2349
Laringite	1841	1923
Tonsillite	2148	2278

Table 4.9: AIC del modello GAM con e senza interazione fra valore ritardato di un giorno dell'ozono e anno.

marginali) come effetto fisso nel modello ad effetti misti che viene stimato nell'algoritmo GMERF. Chiaramente, l'anno e il valore cumulato dell'inquinante non sono stati considerati tra le covariate utilizzate per la foresta casuale. In tabella 4.10 vengono riportati gli effetti in termini di *odds ratio* ottenuti dell'inquinante cumulato nel 2017 e nel 2020, associati ad un incremento di 10 $\mu g/m^3$. Viene inoltre riportato il numero di iterazioni. Si nota che tipicamente l'effetto nel 2020 risulta di minore entità rispetto al 2017. Una tabella riassuntiva analoga mostra i risultati sull'ozono (tab. 4.11), notando tuttavia alcuni risultati incoerenti con quanto ottenuto dai modelli GAM. Secondo i risultati del modello GMERF, infatti, nel 2020 l'impatto dell'ozono sembra ridursi per la maggior parte delle diagnosi. Se dunque sembra confermato il fatto che l'ozono sia un fattore di rischio, non appare chiaro come il suo effetto cambi nei due anni studiati. Può essere importante considerare che l'ozono, ancor più delle polveri sottili, è strettamente dipendente da altri fattori come ad esempio la temperatura. Questo aspetto aumenta la necessità di "aggiustare" l'effetto dell'ozono tenendo conto del confondimento delle altre variabili. Il modello GAM si presenta come un modello multivariato che considera tutte le covariate e permette dunque di depurare l'effetto dell'ozono da quello di altre variabili; al contrario, il modello GMERF alterna lo studio dell'impatto dell'ozono a quello delle altre covariate, separando le due fasi. D'altronde il modello GMERF nasce come algoritmo di classificazione (in generale, di previsione, a seconda della natura della variabile risposta): poiché l'obiettivo dello studio non è predire ma quanto più spiegare, appare plausibile

Diagnosi	Numero iterazioni	OR 2017	OR 2020
Accessi AVR	95	1.004	0.987 *
Apparato respiratorio	27	0.997	0.994
Otite media	15	1.041 ***	1.004
Rinosinusite	11	0.995	0.977 ***
Faringite	43	1.016 *	1.002
Laringite	30	1.013	1.021 *
Tonsillite	60	0.996	1.006

Table 4.10: GMERF: numero di iterazioni, effetti (*odds ratio*) del valore cumulato delle polveri sottili stimati dal modello a effetti misti. Gli asterischi indicano la significatività: *** per $p.value < 0.001$, ** per $0.001 \leq p.value < 0.01$, * per $0.01 \leq p.value < 0.05$.

considerare più attendibili i risultati del modello GAM.

4.3 Conclusioni

L'età è il principale fattore che agisce sulla probabilità di accedere in pronto soccorso per diagnosi legate alle alte vie respiratorie rispetto all'accesso per trauma. In particolare, l'età risulta un fattore di rischio fino ai 50 anni circa per le diagnosi di laringite e otite media, mentre per le altre diagnosi l'effetto diventa protettivo già a partire dai 25 anni circa. Osservando l'effetto parametrico dell'età categorizzata in 4 classi si nota che, considerando come riferimento la classe 0-25 anni, l'effetto diminuisce di entità ad aumentare della classe d'età. Questo è dovuto al fatto che per le diagnosi si registrano pochi accessi di soggetti anziani, rispetto agli accessi per trauma che coinvolgono soggetti di tutte le età. Tra gli allergeni più impattanti sul rischio di accedere per malattia rispetto al trauma si trovano l'alternaria e le betulacee, mentre fra gli inquinanti si trovano l'ozono e le polveri sottili. I risultati vengono confermati anche nel momento in cui viene presa in considerazione la struttura gerarchica dei dati. Per il confronto con l'anno 2017 è stato scelto di concentrarsi sull'effetto della media mobile delle polveri sottili, analogamente a quanto fatto nel capitolo 3, e sul valore ritardato di un giorno dell'ozono. I modelli GAM che includevano l'interazione fra

Diagnosi	Numero iterazioni	OR 2017	OR 2020
Accessi AVR	43	1.013 ***	1.011 ***
Apparato respiratorio	12	1.003	1.0002
Otite media	54	1.006	1.019 ***
Rinosinusite	8	1.005 *	1.007 **
Faringite	16	1.009 *	1.007
Laringite	8	1.014 ***	1.005
Tonsillite	57	1.012 **	1.002

Table 4.11: GMERF: numero di iterazioni, effetti (*odds ratio*) del valore ritardato di un giorno dell'ozono stimati dal modello a effetti misti. Gli asterischi indicano la significatività: *** per $p.value < 0.001$, ** per $0.001 \leq p.value < 0.01$, * per $0.01 \leq p.value < 0.05$.

anno e inquinante sono risultati aventi un adattamento migliore rispetto a quelli senza interazione. Tuttavia, per quanto riguarda le polveri sottili, gli effetti trovati nei due anni non sono apparsi significativamente diversi da 0, mentre l'ozono è emerso come fattore di rischio in modo significativo nel 2020. L'utilizzo del modello GMERF con l'inserimento dell'interazione nella parte del modello ad effetti misti evidenzia alcune diminuzioni significative dal 2017 al 2020 dell'effetto delle polveri sottili mentre, pur identificando l'ozono come fattore di rischio in entrambi gli anni, non appare coerente con quanto emerso dal modello GAM. Si segnala che il modello GMERF potrebbe non essere quello più idoneo per quantificare un effetto in quanto nasce come algoritmo di classificazione e nella stima non tiene conto del possibile confondimento dovuto ad altre covariate.

Capitolo 5

Conclusioni e limiti dello studio

I dati analizzati in questo elaborato sono riferiti a due anni di studio, il 2020 e il 2017. I dati raccolti nel 2020 sono inizialmente stati analizzati separatamente da quelli riferiti al 2017, già studiati in Ottaviano et al. [2022]. Quest'ultimo studio aveva identificato come fattori di rischio l'alternaria sui ricoveri giornalieri totali in pronto soccorso per cause legate alle alte vie respiratorie, e i valori ritardati della concentrazione delle polveri sottili sui ricoveri per rinosinusite, laringite e otite media. Anche nel 2020 l'alternaria presenta un'associazione positiva con alcune diagnosi specifiche, inoltre tra gli allergeni più impattanti sulla salute emergono anche le betulacee. L'ozono ritardato di un giorno e il valore cumulato delle polveri sottili su quattro giorni consecutivi risultano associati positivamente con rinosinusite e faringite. Le analisi sottolineano come sia importante studiare ciascuna diagnosi separatamente, invece di concentrarsi sugli accessi totali, per identificare in modo più specifico i fattori di rischio legati a ciascuna patologia.

Il confronto tra i due anni aveva lo scopo di determinare se diverse condizioni di vita, in questo caso marcatamente cambiate a causa della pandemia, avessero determinato diverse associazioni tra gli inquinanti e i ricoveri in pronto soccorso. Non è facile stabilire a cosa siano dovute le differenze emerse. Nel caso delle polveri sottili, che nel 2020 presentano concentrazioni più basse rispetto al 2017, è plausibile pensare che minori livelli di PM_{10} siano associati ad una manifestazione di minore entità delle malattie legate alle alte vie respiratorie, anche se le numerosità ridotte impedivano

di riscontrare risultati significativi. La concentrazione dell'ozono, invece, è simile nei due anni. L'associazione con l'ozono appare più marcata nel 2020 rispetto al 2017 per le diagnosi di otite e rinosinusite, mentre risulta meno chiaro come l'associazione si modifichi nei due anni per le restanti diagnosi.

Lo studio presenta alcuni limiti. Il primo è la difficoltà di tenere conto del fatto che nel 2020 gli accessi in pronto soccorso potevano essere influenzati non solo dai fattori ambientali ma anche dalla situazione pandemica. Basare il confronto su periodi non affetti da confondimenti di questo tipo aiuterebbe a stabilire con maggiore correttezza le reali differenze degli effetti in periodi diversi. Sarebbe inoltre opportuno considerare le serie temporali su più anni in modo da disporre di una numerosità più elevata e ottenere risultati più robusti.

In ogni caso, lo studio è a supporto delle numerose ricerche che dichiarano l'effetto dannoso dell'inquinamento sulla salute. Se i fattori climatici, come temperatura e umidità, non sono controllabili dall'uomo, è possibile però agire su altri elementi che hanno un'influenza sulla salute pubblica. Oltre a scegliere in maniera accurata le piante arboree da collocare in città in modo da evitare per quanto possibile la presenza di aeroallergeni dannosi, è importante adottare misure per ridurre l'inquinamento atmosferico e sensibilizzare la popolazione nel preferire comportamenti consapevoli e che limitino l'impatto sull'ambiente, anche nell'interesse della propria salute.

Bibliografia

ARPAV. Livelli di concentrazione di polveri fini (pm10). https://www.arpa.veneto.it/arpavinforma/indicatori-ambientali/indicatori_ambientali/atmosfera/qualita-dellaria/livelli-di-concentrazione-di-polveri-fini-pm10/2021, 2021. Ultimo accesso 1 Maggio 2023.

ARPAV. Rapporto 2017. https://www.arpa.veneto.it/temi-ambientali/pollini/file-e-allegati/storico-rapporto-pollini/rapporto-2017, 2022. Ultimo accesso 31 Luglio 2023.

ARPAV. Livelli di concentrazione di biossido di azoto (no2). https://www.arpa.veneto.it/arpavinforma/indicatori-ambientali/indicatori_ambientali/atmosfera/qualita-dellaria/livelli-di-concentrazione-di-biossido-di-azoto-no2/2021, 2022. Ultimo accesso 2 Maggio 2023.

ARPAV. Bollettino livelli di allerta pm10. https://www.arpa.veneto.it/dati-ambientali/bollettini/aria/bollettino-livelli-di-allerta-pm10, 2023. Ultimo accesso 1 Maggio 2023.

ARPAV. A proposito di ozono... https://www.arpa.veneto.it/temi-ambientali/aria/a-proposito-di-ozono, 2023. Ultimo accesso 2 Maggio 2023.

S Blackmore, Jocelyn AJ Steinmann, PP Hoen, and W Punt. Betulaceae and corylaceae. *Review of Palaeobotany and Palynology*, 123(1-2):71–98, 2003.

Jeroen Buters et al. Ambrosia artemisiifolia (ragweed) in germany–current presence, allergological relevance and containment procedures. *Allergo Journal International*, 24:108–120, 2015.

CNR. Misurare la pioggia. https://polaris.irpi.cnr.it/misurare-la-pioggia/, 2021. Ultimo accesso 8 Maggio 2023.

Gennaro D'Amato et al. Urban air pollution and climate change as environmental risk factors of

respiratory allergy: an update. *Journal of Investigational Allergology and Clinical Immunology*, 20(2):95–102, 2010.

Letty A de Weger et al. The long distance transport of airborne ambrosia pollen to the uk and the netherlands from central and south europe. *International journal of biometeorology*, 60: 1829–1839, 2016.

European Environmental Agency. Air quality in europe—2020 report. *European Environmental Agency*, 2020.

Vicki Ann Funk et al. Classification of compositae. *Systematics, evolution, and biogeography of Compositae*, 2009.

H García-Mozo. Poaceae pollen as the leading aeroallergen worldwide: A review. *Allergy*, 72(12): 1849–1858, 2017.

Fabrizio Giostra et al. Impact of covid-19 pandemic and lockdown on emergency room access in northern and central italy. *Emerg Care J*, 17(2):9705, 2021.

Peter J Green and Bernard W Silverman. *Nonparametric regression and generalized linear models: a roughness penalty approach*. Crc Press, 1994.

Trevor J Hastie and Robert J Tibshirani. *Generalized additive models*, volume 43. CRC press, 1990.

C L Jones. Environmental and clinical mould spore risk thresholds. *Journal of Bacteriology & Mycology: Open access*, 11(1):44–48, 2023.

Ute Latza, Silke Gerdes, and Xaver Baur. Effects of nitrogen dioxide on human health: systematic review of experimental and epidemiological studies conducted between 2002 and 2006. *International journal of hygiene and environmental health*, 212(3):271–287, 2009.

Legambiente. Mal'aria di città, 2022.

Legambiente. Mal'aria di città, 2023.

Jakub Nowosad et al. Forecasting model of corylus, alnus, and betula pollen concentration levels using spatiotemporal correlation properties of pollen count. *Aerobiologia*, 32:453–468, 2016.

Hanna Ojrzyńska et al. The influence of atmospheric circulation conditions on betula and alnus pollen concentrations in wrocław, poland. *Aerobiologia*, 36:261–276, 2020.

Organizzazione Mondiale della Sanità. Linee guida globali oms sulla qualità dell'aria: particolato (pm2, 5 e pm10), ozono, biossido di azoto, anidride solforosa e monossido di carbonio: sintesi. Technical report, Organizzazione Mondiale della Sanità. Ufficio Regionale per l'Europa, 2022.

Claudio Ortolani et al. Allergenicità delle piante arboree e arbustive destinate al verde urbano italiano. revisione sistematica e raccomandazioni basate sull'evidenza. *Giornale Europeo di Aerobiologia Medicina Ambientale e Infezioni Aerotrasmesse*, 11, 2015.

Giancarlo Ottaviano et al. The impact of air pollution and aeroallergens levels on upper airway acute diseases at urban scale. *International Journal of Environmental Research*, 16(4):42, 2022.

Massimo Pellagatti et al. Generalized mixed-effects random forest: A flexible approach to predict university student dropout. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 14(3):241–257, 2021.

Krystyna Piotrowska-Weryszko and Elżbieta Weryszko-Chmielewska. Plant pollen content in the air of lublin (central-eastern poland) and risk of pollen allergy. *Annals of Agricultural and Environmental Medicine*, 21(4), 2014.

Alberto Pivato et al. Long time series analysis of air quality data in the veneto region (northern italy) to support environmental policies. *Atmospheric Environment*, 298:119610, 2023.

Kathrin Reinmuth-Selzle et al. Chemical modification by peroxynitrite enhances tlr4 activation of the grass pollen allergen phlp5. *Frontiers in Allergy*, 4, 2023.

Rete Italiana di Monitoraggio Aerobiologico. Pollnet - bollettini dei pollini e monitoraggio aerobiologico. http://www.pollnet.it/valori_di_riferimento_it.asp, 2022. Ultimo accesso 30 Aprile 2023.

SNPAmbiente. Biossido di azoto (no2). https://www.snpambiente.it/temi/biossido-di-azoto/, 2020. Ultimo accesso 2 Maggio 2023.

SNPAmbiente. La qualità dell'aria in europa. https://www.snpambiente.it/2020/12/01/la-qualita-dellaria-in-europa-2/, 2020. Ultimo accesso 4 Luglio 2023.

Nadine Steckling-Muschack et al. A systematic review of threshold values of pollen concentrations for symptoms of allergy. *Aerobiologia*, 37(3):395–424, 2021.

Anna Tosi et al. Time lag between ambrosia sensitisation and ambrosia allergy. *Swiss medical weekly*, 141(3940):w13253–w13253, 2011.

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

Appendice A

Codice R

A.1 Selzione stepwise nel modello GAM

```
# response = nome della variabile risposta di conteggio
# covariate = vettore di nomi delle covariate
# dati = dataset contenente response e covariate
gam.stepwise <- function(response, covariate, dati){

  # modello con sola intercetta
  model_f <- paste(response, '~ 1')
  m0 <- mgcv::gam(as.formula(model_f), family = poisson, data = dati, method = 'REML')
  best_aic <- AIC(m0)

  # impostazione valori inziali
  best_var <- NULL
  insert <- NULL
  remain <- covariate
  stop <- 0

  # selezione stepwise
  while (stop == 0) {

    cat('Current model: ', model_f, ' ( AIC = ', best_aic, ')', '\n')
    best_var <- NULL
```

```
    for(variable in remain){

      # modello con le covariate gia' selezionate piu' nuova variabile da testare
      f_try <- paste(response, "~",
                     paste0("s(", c(insert, variable), ")", collapse = " + "), sep = " ")
      m_try <- mgcv::gam(as.formula(f_try), family = poisson, data = dati, method = 'REML')
      aic_try <- AIC(m_try)

      cat(variable, ': AIC = ', aic_try, '\n')

      if(aic_try < best_aic){
        best_aic <- aic_try
        best_var <- variable
      }
    }

    if(is.null(best_var)){
      cat('\nNo more covariate added.\n')
      stop = 1 # fine della ricerca
    } else {
      insert <- c(insert, best_var) # variabili inserite nel modello
      remain <- setdiff(covariate, insert) # variabili non inserite e che
                                           # potrebbero apportare un miglioramento
    }

    model_f <- paste(response, "~",
                     paste0("s(", c(insert), ")", collapse = " + "), sep = " ")
    cat('\n')
  }

  return(model_f)
}
```

A.2 Esempio modello GAMM su dati simulati

```
library(itsadug) # per la funzione plot_smooth

# simulazione dati
x <- runif(n = 50, min = 0, max = 3)
y1 <- -x^3 + 3*x^2 + rnorm(50, sd = 0.3)
y2 <- 1 - 0.8*x^3 + 2*x^2 + rnorm(50, sd = 0.3)
y3 <- 0.5 - 0.8*x^3 + 1.9*x^2 + rnorm(50, sd = 0.3)
y4 <- 0.5*x^2 + rnorm(50, sd = 0.3)

sim <- as.data.frame(cbind(x = rep(x, 4), y = c(y1, y2, y3, y4),
group = c(rep(1,50),rep(2,50),rep(3,50),rep(4,50))))
sim$group <- as.factor(sim$group)

# modello intercetta casuale
gam_intercept <- gam(y ~ s(x) + s(group, bs = 're'), data = sim, method = 'REML')
summary(gam_intercept)
plot_smooth(gam_intercept, view = "x", rm.ranef = FALSE, cond = list(group = "1"),
    main = "Random intercept", col = "orange", ylim = c(-2.5, 4.2), h0=NULL)
plot_smooth(gam_intercept, view = "x", rm.ranef = FALSE, cond = list(group = "2"),
    add = TRUE, col = "red")
plot_smooth(gam_intercept, view = "x", rm.ranef = FALSE, cond = list(group = "3"),
    add = TRUE, col = "purple")
plot_smooth(gam_intercept, view = "x", rm.ranef = FALSE, cond = list(group = "4"),
    add = TRUE, col = "turquoise")

# modello pendenza casuale
gam_slope <- gam(y ~ s(x) + s(x, group, bs = 're'), data = sim, method = 'REML')
summary(gam_slope)
plot_smooth(gam_slope, view = "x", rm.ranef = FALSE, cond = list(group = "1"),
    main = "Random slope", col = "orange", ylim = c(-2.5, 4.2), h0=NULL)
plot_smooth(gam_slope, view = "x", rm.ranef = FALSE, cond = list(group = "2"),
    add = TRUE, col = "red")
plot_smooth(gam_slope, view = "x", rm.ranef = FALSE, cond = list(group = "3"),
    add = TRUE, col = "purple")
plot_smooth(gam_slope, view = "x", rm.ranef = FALSE, cond = list(group = "4"),
    add = TRUE, col = "turquoise")
```

```r
# modello intercetta casuale + pendenza casuale
gam_int_slope <- gam(y ~ s(x) + s(group, bs = 're') + s(x, group, bs = 're'), data = sim,
method = 'REML')
summary(gam_int_slope)
plot_smooth(gam_int_slope, view = "x", rm.ranef = FALSE, cond = list(group = "1"),
    main = "Random intercept + random slope", col = "orange", ylim = c(-2.5, 4.2), h0=NULL)
plot_smooth(gam_int_slope, view = "x", rm.ranef = FALSE, cond = list(group = "2"),
    add = TRUE, col = "red")
plot_smooth(gam_int_slope, view = "x", rm.ranef = FALSE, cond = list(group = "3"),
    add = TRUE, col = "purple")
plot_smooth(gam_int_slope, view = "x", rm.ranef = FALSE, cond = list(group = "4"),
    add = TRUE, col = "turquoise")

# modello random smooth
gam_smooth <- gam(y ~ s(x) + s(x, group, bs = 'fs', m = 1), data = sim, method = 'REML')
summary(gam_smooth)
plot_smooth(gam_smooth, view = "x", rm.ranef = FALSE, cond = list(group = "1"),
    main = "Random smooth", col = "orange", ylim = c(-2, 4.2), h0=NULL)
plot_smooth(gam_smooth, view = "x", rm.ranef = FALSE, cond = list(group = "2"),
    add = TRUE, col = "red")
plot_smooth(gam_smooth, view = "x", rm.ranef = FALSE, cond = list(group = "3"),
    add = TRUE, col = "purple")
plot_smooth(gam_smooth, view = "x", rm.ranef = FALSE, cond = list(group = "4"),
    add = TRUE, col = "turquoise")
```

A.3 Algoritmo GMERF

```r
gmerf <- function(cov = cov, # nomi covariate
                  id = id, # nome variabile di gruppo
                  data = data,
                  znam = NULL, # nomi variabili effetti casuali
                  toll = 0.01, itmax = 30, mtry = 5, ntree = 300){
```

```r
set.seed(3782)
data <- data %>% drop_na()

# matrice llik
llik_mat <- matrix(NA, ncol = 2, nrow = itmax)
llik_mat[,1] <- c(1:itmax)

# stimo glm con tutte le covariate ed estraggo predittore lineare
f_glm <- as.formula(paste('y ~', paste(cov, collapse = ' + ')))
m_glm <- glm(formula = f_glm, family = binomial, data = data)
eta <- m_glm$linear.predictors

it <- 1 # iterazione
conv <- FALSE # convergenza
llk_0 <- 0

#  variabile dipendente iniziale del RF
# (all'inizio gli effetti casuali sono assunti pari a 0)
target <- eta

# formula generale del RF
f_rf <- as.formula(paste('target ~', paste(cov, collapse = ' + ')))

# formula generale del LMM
if(is.null(znam)){ # solo intercetta casuale
  f_lmm <- as.formula(paste('reff ~ - 1 + (1 | ', id, ')'))
  # per valutare interazione fra anno e pm10
  # f_lmm <- as.formula(paste('reff ~ anno + I(pm10_mm/10):anno + (1 | ', id, ')'))
  # per valutare interazione fra anno e ozono
  # f_lmm <- as.formula(paste('reff ~ anno + I(ozone_1/10):anno + (1 | ', id, ')'))
}else{ # intercetta + pendenza casuale
  f_lmm <- as.formula(paste('reff ~ - 1 + (1 + ', paste(znam, collapse = '+'), '| ', id, ')'))
}

while(it <= itmax & conv == FALSE){

  # stimo RF usando target come risposta con tutte le covariate
  m_rf = randomForest(formula = f_rf, data = data, mtry = mtry, ntree = ntree,
```

```
                        do.trace=50, importance = T)

    # ottengo risposta su cui stimare LMM
    reff <- eta - m_rf$predicted
    m_lmm <- lmer(formula = f_lmm, data = data)

    # definisco nuova risposta su cui stimare RF
    target <- eta - predict(m_lmm)

    # convergenza
    llk <- as.numeric(logLik(m_lmm))
    llik_mat[it,2] <- llk
    tr <- abs(llk - llk_0)
    if(tr < toll){conv <- TRUE}else{
      print(paste('Iteration ', it, ': ', llk))
      llk_0 <- llk
      it <- it +1
      }
  }

  if(conv == FALSE){print('Warning: maximum number of iterations reached.')
    return(list(llk = llik_mat))}else{
    result <- list(forest = m_rf, lmm = m_lmm, llk = llik_mat)
    return(result)
  }
}
```