Progetto PULVIRUS

OBIETTIVO 1 - Analisi degli effetti delle misure di distanziamento fisico durante il periodo della pandemia da covid 19: cosa dicono le stazioni di monitoraggio italiane.

ATTIVITÀ 1.2 – Ricognizione della letteratura scientifica sui metodi di normalizzazione meteorologica e sui metodi di "intervention analysis" applicati a serie storiche di dati di qualità dell'aria.

Bozza di relazione tecnica: ver. 1.0

Data: 21/12/2021

GRUPPO DI LAVORO

ISPRA

Gianluca Leone (Coordinatore), Mariacarmela Cusano

ENEA

Ilaria D'Elia, Maria Gabriella Villani

Arpa Emilia Romagna

Fabiana Scotto

Arpa Lombardia

Andrea Algieri

SOMMARIO

INTRODUZIONE

La presente relazione sintetizza i contenuti tecnico scientifici desunti dall'analisi della letteratura, aggiornata al mese di ottobre 2020, inerente alle tre fondamentali famiglie di modelli statistici: i modelli additivi generalizzati (GAM), i modelli autoregressivi multivariati, i modelli *Random Forest*.

Tali modelli sono stati impiegati, in molti dei lavori analizzati, per cercare di stimare l'efficacia di misure finalizzate al miglioramento della qualità dell'aria, pertanto rappresentano un punto imprescindibile anche per la valutazione degli effetti che il *lockdown* della primavera del 2020 ha indotto sull'inquinamento atmosferico in Italia. Gli articoli scientifici sono stati analizzati e sintetizzati singolarmente mediante una scheda descrittiva, con l'intento di fornire, nell'ambito del gruppo di lavoro del progetto Pulvirus, una comune base conoscitiva utile per definire i passi più opportuni per l'implementazione dei modelli considerati.

Si introduce inoltre una prima analisi sul tema dei test multipli nell'analisi spazio temporale dei dati ambientali.

1 MODELLI GAM

1.1 Introduzione

I modelli statistici additivi generalizzati (Generalised Additive Models, GAM), adoperando funzioni di lisciamento ("smoothing"), consentono di valutare interazioni di tipo non lineare tra le covariate e la variabile risposta anche nel caso, molto frequente, in cui non ci sia una conoscenza a priori del tipo di legame funzionale.

La forma funzionale standard del modello GAM è così definita (Wood, 2017):

$g(\mu i)=A i\gamma+\Sigma f j(x j i)\ j$, yi ~ EF(μi, φ)

con:

yi = variabile risposta

μi ≡ E(yi)= valore atteso di yi

yi ~ EF(μi, φ)= distribuzione esponenziale di yi

$A i\gamma$ = *ith* riga della matrice dei parametri del modello con il suo corrispondente vettore

$f j(x j i)$ = funzione di smoothing per le j covariate

L'utilizzo delle funzioni di smoothing in luogo delle funzioni deterministiche basate, nei modelli lineari, sulla stima dei parametri di regressione, ha prodotto ottimi risultati nell'analisi dei sistemi ecologici complessi (Zuur et alii, 2009 e Zuur, 2012). L'impiego di spline quali funzioni di "lisciamento" consente sia di riprodurre l'andamento globale del contributo che la covariata fornisce alla variabile risposta, sia di approssimare al meglio eventuali andamenti locali in particolari intervalli del dominio di esistenza della variabile esplicativa. Tale positiva caratteristica è legata al fatto che le funzioni spline interpolano l'insieme dei punti (yi, xji), suddividendo il dominio di esistenza in intervalli più piccoli, in ognuno dei quali viene impiegato un polinomio, di grado basso, che assicuri, in modo liscio, la continuità nei nodi degli stessi intervalli. È evidente da quanto sopra descritto come la definizione del grado ottimale di

lisciamento rappresenti un aspetto critico nell'uso delle spline: un numero eccessivo di intervalli può portare all'overfitting del modello che implica scarse performance nella fase previsionale. Al contrario un numero troppo basso di intervalli può portare al risultato diametralmente opposto (underfitting) con la conseguente incapacità del modello di seguire l'andamento dei dati in tutto il dominio di esistenza.

1.2 Scheda analisi bibliografia GAM n°1

Barmpadimos, I., Hueglin, C., Keller, J., Henne, S., & Prévôt, A. S. H. (2011). Influence of meteorology on PM 10 trends and variability in Switzerland from 1991 to 2008. *Atmospheric* Chemistry *and Physics, 11*, 1813–1835. https://doi.org/10.5194/acp-11-1813-2011

Metodo statistico per la normalizzazione e stima del contributo

a. Dati di input

i. Dati meteo

(§2 pag. 1815- 1816)

Quali principali dati di input meteorologici (wind speed, wind direction, precipitation, global radiation, net radiation, air temperature and relative humidity), sono stati utilizzati i parametri rilevati presso le stazioni meteorologiche della rete di MeteoSwiss e NABEL. I dati meteo sono stati controllati per verificarne l'attendibilità (presenza di outlier, coerenza tra le variabili e tra punti di misura). Dai dati di base sono state calcolate le medie su base giornaliera e su tre periodi della giornata, mattina (6-12), pomeriggio (12-18) e sera (18-24). Partendo da dati mediati su intervalli a 10 minuti è stata ottenuta la raffica oraria (wind gust) e da quelli orari la raffica di vento giornaliera.

Per la valutazione dell'altezza del PBL è stata utilizzata la variabile proxy del gradiente verticale della temperatura potenziale, considerando la differenza tra la temperatura potenziale (approssimata) rilevata tra due stazioni vicine ma ubicate in siti caratterizzati da un forte differenza di quota (Ordonez et al., 2005). In aggiunta, sempre per la stima

dell'altezza del PBL, è stato calcolato il parametro CAPE (convective available potential energy) partendo dai dati rilevati in una specifica stazione rurale.

Sono stati inoltre considerati alcuni parametrici meteorologici sinottici:

- North Atlantic Oscillation (NAO) index su base mensile, calcolato per singola stazione (Jones et al., 1997).
- La variabile categoriale "synoptic group", così come calcolata in (Wanner et al., 1998), per capire l'influenza delle condizioni meteo generali sui livelli di PM_{10}.
- La variabile categoriale "Fronte della perturbazione" e quella scalare "numero di giorni dal fronte della perturbazione".
- "Yesterday precipitation" cioè l'altezza di pioggia del giorno precedente.

Sono state incluse nel modello statistico soltanto le variabili che, per ogni stazione e per ogni stagione, hanno garantito almeno il 70% di dati validi.

A pag. 1827 viene consigliata una serie storica minima di dati pari a 7 anni.

Variabili meteo	Risoluzione spaziale	Risoluzione temporale	Serie storica minima	Fonti dati	Criteri di esclusione	Note
Wind speed	Puntuale	d**	15 anni	MeteoSwiss e NABEL	Dati validi a livello stagionale > 70%	
Wind gust	Puntuale	d**	15 anni	MeteoSwiss e NABEL	Almeno il 70 % di dati a livello stagionale.	
Surface pressure	Puntuale	d**	15 anni	MeteoSwiss e NABEL	Almeno il 70 % di dati a livello stagionale.	
Global irradiance	Puntuale	d**	15 anni	MeteoSwiss e NABEL	Almeno il 70 % di dati a livello stagionale.	
Net irradiance	Puntuale	d**	15 anni	MeteoSwiss e NABEL	Almeno il 70 % di dati a livello stagionale.	
Daily precipitation	Puntuale	d**	15 anni	MeteoSwiss e NABEL	Almeno il 70 % di dati a livello stagionale.	
Relative humidity	Puntuale	d**	15 anni	MeteoSwiss e NABEL	Almeno il 70 % di dati a livello stagionale.	
Temperature	Puntuale	d**	15 anni	MeteoSwiss e NABEL	Almeno il 70 % di dati a livello stagionale.	

Water vapor mixing ratio	Puntuale	d**	15 anni	MeteoSwiss e NABEL	Almeno il 70 % di dati a livello stagionale.	
Daily sunshine duration	Puntuale	d**	15 anni	MeteoSwiss e NABEL	Almeno il 70 % di dati a livello stagionale.	
Convective available potential energy	Puntuale	d**	15 anni	MeteoSwiss e NABEL	Almeno il 70 % di dati a livello stagionale.	
Afternoon global irradiance	Puntuale	d**	15 anni	MeteoSwiss e NABEL	Almeno il 70 % di dati a livello stagionale.	
Afternoon temperature	Puntuale	d**	15 anni	MeteoSwiss e NABEL	Almeno il 70 % di dati a livello stagionale.	
Afternoon mixing ratio	Puntuale	d**	15 anni	MeteoSwiss e NABEL	Almeno il 70 % di dati a livello stagionale.	
Afternoon wind speed	Puntuale	d**	15 anni	MeteoSwiss e NABEL	Almeno il 70 % di dati a livello stagionale.	
Afternoon sunshine duration	Puntuale	d**	15 anni	MeteoSwiss e NABEL	Almeno il 70 % di dati a livello stagionale.	
Lightning great distance	n.d	d**	15 anni	MeteoSwiss e NABEL	Almeno il 70 % di dati a livello stagionale.	
Lightning at small distance	n.d	d**	15 anni	MeteoSwiss e NABEL	Almeno il 70 % di dati a livello stagionale.	
Convective boundary layer depth	n.d	d**	15 anni	MeteoSwiss e NABEL	Almeno il 70 % di dati a livello stagionale.	Non descritto il metodo calcolo; approfondire in Ordonez , 2005
NAO index	n.d	d**	15 anni	MeteoSwiss e NABEL	Almeno il 70 % di dati a livello stagionale.	Jones et al., 1997
Amount of precipitation the previous day	n.d	d**	15 anni	MeteoSwiss e NABEL	Almeno il 70 % di dati a livello stagionale.	Non descritto il metodo calcolo; approfondire in Ordonez , 2005
Morning wind speed	n.d	d**	15 anni	MeteoSwiss e NABEL	Almeno il 70 % di dati a livello stagionale.	Non descritto il metodo calcolo; approfondire in Ordonez , 2005
Morning global irradiance	n.d	d**	15 anni	MeteoSwiss e NABEL	Almeno il 70 % di dati a livello stagionale.	Non descritto il metodo calcolo; approfondire in Ordonez , 2005
Morning sunshine duration	n.d	d**	15 anni	MeteoSwiss e NABEL	Almeno il 70 % di dati a livello stagionale.	Non descritto il metodo calcolo; approfondire in Ordonez , 2005
Synoptic group	n.d	d**	15 anni	MeteoSwiss e NABEL	Almeno il 70 % di dati a livello stagionale.	Wanner et al., 1998
Front	n.d	d**	15 anni	MeteoSwiss e NABEL	Almeno il 70 % di dati a livello stagionale.	Non descritto il metodo calcolo; approfondire in Ordonez , 2005

Number of days since front	n.d	d**	15 anni	MeteoSwiss e NABEL	Almeno il 70 % di dati a livello stagionale.	Non descritto il metodo calcolo; approfondire in Ordonez , 2005
Vertical gradient of potential temperature	n.d	d**	15 anni	MeteoSwiss e NABEL	Almeno il 70 % di dati a livello stagionale.	Non descritto il metodo calcolo; approfondire in Ordonez, 2005
Lightning at afternoon and evening	n.d	d**	15 anni	MeteoSwiss e NABEL	Almeno il 70 % di dati a livello stagionale.	Non descritto il metodo calcolo; approfondire in Ordonez, 2005

*nd = non disponibile **d= giornaliero

Il formato dei dati non è specificato nell'articolo.

(§4.1 pag. 1819)
Le variabili esplicative più frequentemente selezionate per stagione sono riportate nella tabella sottostante con l'indicazione del segno della loro relazione con la variabile dipendente (PM_{10}).

Primavera	Estate	Autunno	Inverno	Anno
wind gust (−) CBL depth (−) y. precip. (−) rel. humidity (−) net irradiance (−) daily precip. (−)	Julian day (−) wind gust (−) a. temp. (+)	wind gust (−) CBL depth (−) y. precip. (−) daily precip (−)	wind gust (−) CBL depth (−) temperature (−) y. precip. (−) net irradiance (+) daily precip. (−)	daily precip (−) wind gust (−) Julian day (−) y. precip. (−) d. s. front (+) CBL depth (−) temperature

ii. <u>Dati qualità dell'aria</u>

Nello studio sono stati presi in considerazione 13 punti di misura del PM_{10} nel territorio svizzero.

Inquinante	Risoluzione spaziale	Risoluzione temporale	Serie storica minima	fonti dati	Criteri di esclusione
PM_{10}	puntuale	d*	15 anni	Swiss National Air Pollution Monitoring Network NABEL	**n.d.

*d= giornaliero **nd = non disponibile

iii. Altri dati

Per le stazioni di qualità dell'aria sono state inoltre fornite le metainformazioni sulle coordinate geografiche, la quota sul livello del mare e il tipo di stazione.

Come variabili temporali considerate nel modello sono stati presi in considerazione i parametri:

- Giorno giuliano
- Giorno della settimana.

iv. Analisi criticità e punti di forza.

Le variabili meteo non sono sufficientemente descritte. Si dovrà approfondire sugli altri articoli citati, se necessario.

b. Metodo statistico normalizzazione

§3.1 pag. 1816

Per il modello GAM in generale si rimanda all'introduzione. L'implementazione del modello è stata realizzata attraverso il pacchetto mgcv di R. Sono state utilizzate come funzioni di smoothing le thin regression spline. La selezione delle variabili esplicative significative avviene con una procedura "forward" ciclica in 6 step. La procedura sceglie le variabili significative sulla base dell'Akaike Information Criterion (AIC) e attraverso un ciclo complesso. Viene poi calcolato il VIF per verificare la collinearità tra le covariate. Si continua ad aggiungere variabili fino a quando l'AIC continua a diminuire (§3.2 pag. 1817). Poi si aggiunge il giorno giuliano se non entrato nella selezione delle variabili significative. Infine si verificano gli assunti di base.

La concentrazione del PM_{10} normalizzata è espressa mediante la seguente formula:

$$\ln PM_{10}\ adj = a + s(\text{Julian day}) + \varepsilon$$

Non è disponibile il codice sorgente.

(§4.2 pag. 1825-1827)

Per valutare le prestazioni del modello sono stati impiegati i seguenti indicatori:

- Devianza spiegata
- Errore quadratico medio (MSE)

i. Validazione del modello (§4.4pag. 1830)

Il modello è stato sviluppato utilizzando il 90% delle osservazioni disponibili scelte in maniera casuale. Il restante 10% è stato utilizzato per validarlo. I valori predetti sono stati confrontati con quelli misurati corrispondenti ed è stato calcolato l'indicatore "factor of two" (FAC2) per validare quantitativamente il modello.

ii. Analisi criticità e punti di forza

Si sarebbero potuti utilizzare ulteriori parametri per la valutazione quantitativa del modello.

c. Metodo statistico per la stima del contributo delle misure attuate

i. Descrizione algoritmo (formulazione, pacchetti R utilizzati, disponibilità codice sorgente, presenza descrizione procedure)

(§4.3)

Sono stati stimati i trend per la serie delle misure giornaliere e per quella normalizzata. Anche il trend mediato su tutti i punti di misura è stato calcolato. L'intervallo di confidenza di questo è stato stimato tramite una formula di propagazione dell'errore.

Non c'è una stima dell'effetto di misure specifiche.

1.3 Scheda analisi bibliografia GAM n°2

Carslaw, D. C., & Carslaw, N. (2007). Detecting and characterising small changes in urban nitrogen dioxide concentrations. *Atmospheric Environment, 41*(22), 4723–4733. https://doi.org/10.1016/j.atmosenv.2007.03.034

Metodo statistico per la normalizzazione e stima del contributo

a. Dati di input

i. Dati meteo

(§2.1 pag. 5291)

Sono stati utilizzati i dati meteorologici della stazione meteo aeroportuale. I parametri sono rilevati a 10 m dal suolo. Per l'altezza del PBL non viene specificata la fonte dati, né viene definito se il parametro è calcolato ad uno specifico orario o se vengono considerati i valori estremi giornalieri.

Variabili meteo	Risoluzione spaziale	Risoluzione temporale	Serie storica minima	Fonti dati	Criteri di esclusione	Note
Wind speed	Puntuale	d*	8 anni	Meteorological Office Heathrow Airport	n.d.*	
Wind direction	Puntuale	d*	8 anni	Meteorological Office Heathrow Airport	n.d.*	
Wind speed (componenti u,v)	Puntuale	d*	8 anni	Meteorological Office Heathrow Airport	n.d.*	
Temperature	Puntuale	d*	8 anni	Meteorological Office Heathrow Airport	n.d.*	
Relative humidity	Puntuale	d*	8 anni	Meteorological Office Heathrow Airport	n.d.*	
Daily precipitation	Puntuale	d*	8 anni	Meteorological Office Heathrow Airport	n.d.*	
PBL depth	Puntuale?	d*	8 anni	n.d.*	n.d.*	Non è specificato bene su

						quale base temporale è calcolato il parametro.
Cloud cover	Puntuale	d*	8 anni	Meteorological Office Heathrow Airport	n.d.**	

*d= giornaliero **nd = non disponibile

Il formato dei dati non è specificato nell'articolo.

(§3.2 pg. 5292)

Le variabili esplicative selezionate per stagione sono riportate nella tabella sottostante.

Variabili esplicative significative per tutti gli inquinanti
- Interazione tra componente u e v del vento - Temperatura - Giorno giuliano - Flussi di traffico veicoli leggeri - Flussi di traffico veicoli pesanti - Anno - Concentrazione di ozono (solo per NO_2)

ii. <u>Dati qualità dell'aria</u>

Nello studio è stato preso in considerazione un unico punto di misura nel centro di Londra. Gli inquinanti oggetto di studio sono riportati nella tabella seguente

Inquinanti	Risoluzione spaziale	Risoluzione temporale	Serie storica minima	Fonti dati	Criteri di esclusione
NO_2, NO_x, CO, benzene, butadiene	puntuale	d	8 anni	n.d.**	Le medie giornaliere sono calcolate solo se ci sono 12 h di dati validi nella giornata.

*d= giornaliero **nd = non disponibile

iii. <u>Altri dati</u>

I flussi di traffico relativi ai veicoli leggeri e pesanti sono stati inclusi come variabili esplicative del modello.

Come variabili temporali considerate nel modello sono stati presi in considerazione i parametri:

- Giorno giuliano
- Anno

iv. Analisi criticità e punti di forza.

La variabile esplicativa "PBL" non è sufficientemente descritta, pertanto non si riesce a comprendere come sia stata calcolata e da quale fonte provengano i dati.

b. Metodo statistico normalizzazione

i. Descrizione algoritmo

(§3.1 pag. 5292-5295)

Per il modello GAM in generale si rimanda all'introduzione. L'implementazione del modello è stata realizzata attraverso il pacchetto mgcv di R. Sono state utilizzate come funzioni di smoothing le thin regression spline. La selezione delle variabili esplicative significative avviene con una procedura basata sul parametro *generalised cross validation* (GCV). Se il valore dello score GCV diminuisce allorché viene esclusa una covariata ne consegue che l'omissione di tale termine apporta un beneficio al modello. Infine si verificano gli assunti di base.

Come esempio si riporta la concentrazione del PM_{10} normalizzata espressa mediante la seguente formula:

$$\log(NO_2) = s_1(\text{year}) + s_2(u, v) + s_3(\text{temp}) + s_4(\text{JD}) + s_5(\text{lightvehicles}) + s_6(\text{heavyvehicles}) + s_7(O_3) + \varepsilon$$

Non è disponibile il codice sorgente.

(§4.1 pag. 5292-5295)

Per valutare le prestazioni del modello sono stati impiegati i seguenti indicatori:

- Coefficiente di determinazione R^2
- Autocorrelazione dei residui.

Il modello ha mostrato una leggera, ma non insignificante autocorrelazione dei residui. Utilizzando un generalised additive mixed model (GAMM) è stata dimostrata però la non significatività di tale autocorrelazione (risultati non riportati). A seconda dell'inquinante considerato il coefficiente di determinazione R^2 calcolato varia da 0,81 a 0,90.

ii. Validazione del modello

Non presente.

iii. Analisi criticità e punti di forza

Non è descritta la validazione del modello.

c. Metodo statistico per la stima del contributo delle misure attuate

i. Descrizione algoritmo (formulazione, pacchetti R utilizzati, disponibilità codice sorgente, presenza descrizione procedure)

Sono stati stimati i trend annuali (cfr. Holland et al., 2000) per la serie delle misure giornaliere e per quella normalizzata. L'incertezza è stata stimata con la tecnica *bootstrap.* È illustrato un esempio di utilizzo dei modello GAM in forecast.

1.4 Scheda analisi bibliografia GAM n°3

Carslaw, D. C., Beevers, S. D., & Tate, J. E. (2007). Modelling and assessing trends in traffic-related emissions using a generalised additive modelling approach. *Atmospheric Environment, 41*, 5289–5299. https://doi.org/10.1016/j.atmosenv.2007.02.032

Metodo statistico per la normalizzazione e stima del contributo

a. Dati di input

i. Dati meteo

(§2.1 pag. 4725)

Sono stati utilizzati i dati meteorologici di una stazione meteo aeroportuale a 25 km da Londra. I dati, riportati nella tabella sottostante, sono stati aggregati su base giornaliera.

Variabili meteo	Risoluzione spaziale	Risoluzione temporale	Serie storica minima	Fonti dati	Criteri di esclusione
Wind speed	Puntuale	d*	6 anni	Meteorological Office Heathrow Airport	n.d.**
Wind direction	Puntuale	d*	6 anni	Meteorological Office Heathrow Airport	n.d.**
Wind speed (componenti u,v)	Puntuale	d*	6 anni	Meteorological Office Heathrow Airport	n.d.**
Temperature	Puntuale	d*	6 anni	Meteorological Office Heathrow Airport	n.d.**

*d= giornaliero **nd = non disponibile

Il formato dei dati non è specificato nell'articolo.

(§2.3 pag. 4726)

Le variabili esplicative selezionate sono riportate nella tabella sottostante.

Variabili esplicative significative per tutti gli inquinanti
- Interazione tra componente u e v del vento - Temperatura - Concentrazione di NO_x stazione traffico - Concentrazione di NO_2 stazione di fondo - Concentrazione di ozono stazione di fondo

ii. Dati qualità dell'aria

(§2.1 pag. 4725)

Nello studio sono stati presi in considerazione 20 punti di misura di tipo traffico nel centro di Londra. L'inquinante oggetto di studio è il biossido di azoto (NO_2). Per l'elaborazione del modello sono stati utilizzati anche i dati O_3 e NO_2 rilevati nella stazione di fondo più vicina all'area di studio. Si è inoltre fatto uso dei livelli di NO_x rilevati nelle medesime 20 stazioni di traffico sopra menzionate.

Inquinanti	Risoluzione spaziale	Risoluzione temporale	Serie storica minima	Fonti dati	Criteri di esclusione
NO_2	Puntuale	d*	6 anni	London Air Quality Network (LAQN)	Le medie giornaliere sono calcolate solo se ci sono 12 h di dati validi nella giornata.

*d= giornaliero

iii. Altri dati

Le uniche metainformazioni utilizzate si riferiscono al tipo di stazione.

iv. Analisi criticità e punti di forza.

Sono considerate poche variabili meteorologiche. L'utilizzo dei livelli di concentrazione in aria di altri inquinanti come covariate non rende possibile la normalizzazione meteorologica.

b. Metodo statistico normalizzazione

i. Descrizione algoritmo

§2.3 pag. 4726

Per il modello GAM in generale si rimanda all'introduzione. L'implementazione del modello è stata realizzata attraverso il pacchetto mgcv di R. Sono state utilizzate come funzioni di smoothing le penalised regression spline. Nello studio è stato applicato l'approccio di Wood & Augustine (2002) per selezionare le variabili significative e determinare le funzioni di smoothing ottimali.

Si riporta di seguito la formulazione del modello:

$$\log(NO_2) = s1(NO_{2fondo}) + s2(u, v) + s3(O_{3fondo}) + s4(NOx) + \varepsilon$$

Non è disponibile il codice sorgente. Il modello è stato implementato tramite R.

(§3 pag. 4728)

Per valutare le prestazioni del modello sono stati impiegati i seguenti indicatori:

- Coefficiente di determinazione R^2

In funzione del punto di misura considerato il coefficiente di determinazione R^2 calcolato varia da 0,74 a 0,97. Sono stati verificati gli assunti di base del modello.

ii. Validazione del modello

Non presente.

iii. Analisi criticità e punti di forza

Non è descritta la validazione del modello.

c. Metodo statistico per la stima del contributo delle misure attuate

i. Descrizione algoritmo

Mediante il pacchetto R *strucchange s*ono stati stimati i cambiamenti strutturali nelle serie dati. Le serie temporali sono state segmentate e per ogni segmento è stata implementata una regressione lineare. Quindi è stata valutata la significatività dell'ipotesi che i coefficienti dei regressori rimangano invariati da un segmento all'altro. La s*ignificatività di* tale ipotesi è stata valutata con un test basato su F-statistics su una finestra mobile di 365 giorni (i residui del modello segmentato sono stati confrontati con quelli della serie completa).

Non è disponibile il codice sorgente.

Considerare di leggere anche *Wood, S.N., Augustin, N.H. (2002). GAMs with integrated model selection using penalized regression splines and applications to environmental modelling. Ecological Modelling 157 (2–3), 157–177.*

1.5 Scheda analisi bibliografia GAM n°4

Ordóñez, C., Garrido-Perez, J. M., & García-Herrera, R. (2020). Early spring near-surface ozone in Europe during the COVID-19 shutdown: Meteorological effects outweigh emission changes. *Science of the Total Environment, 747(December 2019). https://doi.org/10.1016/j.scitotenv.2020.141322*

Metodo statistico per la normalizzazione e stima del contributo

L'articolo descrive l'effetto delle misure di lockdown sulla qualità dell'aria (O_3 e NO_2) nei mesi di marzo e aprile 2020. Il modello GAM viene sviluppato per il periodo 2015-2019 e utilizzato in forecast per marzo e aprile 2020.

a. Dati di input

i. Dati meteo

(§2.1 pag.747)

Le variabili meteo sono state estratte da ERA5 reanalysis.

Variabili meteo	Risoluzione spaziale	Risoluzione temporale	Serie storica minima	Fonti dati	Criteri di esclusione
Daily maximum air temperature at 2 m (T2max)	0,75°x0,75°	d	6 anni	ECMWF	**n.d.
Mean fields of the zonal (U10) and meridional (V10) wind components at 10 m	0,75°x0,75°	d	6 anni	ECMWF	**n.d.
500 hPa geopotential height (Z500)	0,75°x0,75°	d	6 anni	ECMWF	**n.d.
2-m specific humidity (q)	0,75°x0,75°	d	6 anni	ECMWF	**n.d.
downward solar radiation flux (SR)	0,75°x0,75°	d	6 anni	ECMWF	**n.d.
accumulated precipitation (Prec)	0,75°x0,75°	d	6 anni	ECMWF	**n.d.

*d= giornaliero **nd = non disponibile

I dati sono in formato netcdf.

(§2.3 pag. 4726)

Anche se non ben specificato, sembrerebbe che tutte le variabili meteorologiche di input siano risultate significative nel modello finale.

ii. Dati qualità dell'aria

(§2.1 pag.747)

Nello studio sono presi in considerazione oltre 1300 punti di misura per O_3 e NO_2 i cui dati di monitoraggio sono stati estratti dal DB AirBase dell'Agenzia Europea dell'Ambiente. Lo studio considera solo le stazioni di fondo.

Inquinanti	Risoluzione spaziale	Risoluzione temporale	Serie storica minima	Fonti dati	Criteri di esclusione
NO_2	puntuale	*d(max 1h)	6 anni	AirBase (EEA)	Dati validi > 75% su tutto il periodo di studio
O_3	puntuale	*d(max 8h media mobile)	6 anni	AirBase (EEA)	Dati validi > 75% su tutto il periodo di studio

*d= giornaliero

iii. Altri dati

In aggiunta alle variabili meteo è stata considerata anche la variabile categoriale "working/non-working day".

Le uniche metainformazioni utilizzate si riferiscono al tipo di stazione di monitoraggio della qualità dell'aria.

iv. Analisi criticità e punti di forza.

Si utilizzano dati ERA5 e AirBase come nel progetto Pulvirus.

b. Metodo statistico normalizzazione

i. Descrizione algoritmo

(§2.3 pag. 4726)

Per il modello GAM in generale si rimanda all'introduzione. L'implementazione del modello è stata realizzata attraverso il pacchetto pyGAM in linguaggio Python. Il modello è selezionato considerando le serie dati 2015-2019 soltanto nei mesi interessati dalle misure di Lockdown.

Non è disponibile il codice sorgente.

(§3 pag. 4728)

Per valutare le prestazioni del modello è stato impiegato il seguente indicatore:

- Devianza spiegata

La devianza spiegata mediana è pari al 48% per il biossido di azoto e al 60% per l'ozono.

Sono stati verificati gli assunti di base del modello.

ii. Validazione del modello

Non presente.

iii. Analisi criticità e punti di forza

Non è descritta la validazione del modello: Il valore della devianza spiegata sembra essere inferiore a quello riportato in altri lavori scientifici.

c. Metodo statistico per la stima del contributo delle misure attuate

i. Descrizione algoritmo

Il modello GAM selezionato è stato impiegato in modalità forecast per prevedere i livelli dei due inquinanti nel periodo del *lockdown* (marzo-aprile 2020). A tali valori previsionali

stimati per tutti i punti di misura, sono stati sottratti i livelli misurati in modo tale da poter così valutare l'effetto della fase di *lockdown*. Per l'ozono il contributo dovuto alla meteorologia è stato approfondito applicando il modello GAM con dati meteo mediati su 5 anni (2015 – 2019).

Non è disponibile il codice sorgente.

1.6 Sintesi metodo modello GAM

Dall'analisi della letteratura scientifica esposta nei precedenti paragrafi emergono due direttrici fondamentali di lavoro per l'applicazione dei modelli GAM al fine di valutare gli effetti delle misure di lockdown. La prima, riconducibile a quanto proposto negli articoli di Barmpadimos e di Carslaw, contempla l'utilizzo dei modelli GAM per la ricostruzione delle serie storiche pluriennali degli inquinanti, stazione per stazione, normalizzate rispetto alle variabili meteorologiche. Dall'analisi dei trend della serie normalizzata mediante il pacchetto R *strucchange* è possibile individuare i cambiamenti strutturali nei dati anche attraverso l'identificazione dei *breakpoints* della serie.

L'altro metodo potenzialmente applicabile si fonda su quanto proposto nell'articolo Ordonez. Il modello GAM in questo caso è impiegato prendendo a riferimento soltanto il segmento temporale dell'anno in cui si vogliono stimare gli effetti sulla qualità dell'aria di un determinata politica o misura. La selezione delle variabili significative del modello viene condotta prendendo in considerazione lo stesso segmento temporale sul quale si vogliono effettuare le previsioni per diversi anni antecedenti a quello di analisi; definito in tal modo il modello, si passa poi alla fase previsionale di forecast per il periodo di studio. L'effetto delle misure di *distanziamento sociale* è pertanto stimato sottraendo le previsioni del modello con i dati rilevati dalle stazioni di monitoraggio della qualità dell'aria nell'omologo periodo.

2 MODELLI DI AUTOREGRESSIONE MULTIVARIATA

Sulla base di [Penny and Harrison, 2006], considerata una serie temporale univariata, le sue misurazioni consecutive contengono informazioni sul processo che l'ha generata. Un tentativo di descrivere questo ordine sottostante può essere ottenuto modellando il valore corrente della variabile come una somma lineare ponderata dei suoi valori precedenti. Questo è un processo autoregressivo (AR) ed è un approccio molto semplice, tuttavia efficace, alla caratterizzazione delle serie temporali [Chatfield, 1996]. L'ordine del modello è il numero di osservazioni precedenti utilizzate e i pesi caratterizzano le serie temporali. I modelli autoregressivi multivariati estendono questo approccio a più serie temporali in modo che il vettore dei valori attuali di tutte le variabili sia modellato come una somma lineare di attività precedenti.

Si consideri un numero -*d*- di serie temporali generate da *d*-variabili all'interno di un sistema dove *m* è l'ordine del modello. Un modello MAR (m) prevede il valore successivo in una serie temporale *d*-dimensionale, y_n, come combinazione lineare dei *m* valori vettoriali precedenti:

$$y_n = \sum_{i=1}^{m} y_{n-1} A(i) + e_n$$

dove $y_n = [y_n(1), y_n(2), ..., y_n(d)]$ è l'ennesimo campione di una serie temporale d-dimensionale, ogni A (i) è una matrice di coefficienti (pesi) ed $e_n = [e_n(1), e_n(2), ..., e_n(d)]$ è il rumore gaussiano additivo con media zero e covarianza R. Si noti che si è assunto che la media dei dati sia stata sottratta dal tempo serie.

Il modello può essere scritto nella forma standard di un modello di regressione lineare multivariata come segue:

$$y_n = x_n W + e_n$$

dove $x_n = [y_{n-1}, y_{n-2}, ..., y_{n-m}]$ sono le m serie storiche multivariate dei precedenti campioni e W è una matrice (m × d) -per-d dei coefficienti MAR (pesi). Perciò ci sono un totale di k = m × d × d coefficienti MAR.

2.1 Scheda analisi bibliografia Autoregr. n°1

Xiang J., Austin E., Gould T., Larson T., Shirai J., Liu Y., Marshall J., Seto E. (2020). Impacts of the COVID-19 responses on traffic-related air pollution in a Northwestern US city. *Sci. Tot. Environ*. *747*, 141325. https://doi.org/10.1016/j.scitotenv.2020.141325

Metodo statistico per la normalizzazione e stima del contributo

a. Dati di input

i. Dati meteo

I dati di input meteorologici sono rappresentati nella tabella che segue.

Variabile meteo	Tipo di variabile	Risoluzione spaziale	Risoluzione temporale	Serie storica minima	Fonte dati
W-speed	numerica	puntuale	1 h	17.02.2020 31.05.2020	Seattle 10th &Weller monitoring stations WAQA system
W-dir	categoriale	puntuale	1 h	vedi sopra	vedi sopra
Temperature	numerica	puntuale	1 h	vedi sopra	vedi sopra
RH		puntuale	1 h	vedi sopra	BFI station (7.5 km south) Washingtom state automated surface observing system network
Precipitazione	numerica	puntuale	1 h		Vedi sopra

Alcune note sui dati:

- Il formato dei dati non è disponibile.
- Il periodo temporale totale utilizzato è suddiviso in due intervalli:
  - 17.02.2020 -> 23.03.2020
  - 23.03.2020 -> 31.05.2020
- I dati provengono da un'unica stazione meteo e si sono rilevati due diversi regimi meteorologici.

ii. Dati qualità dell'aria delle stazioni di monitoraggio

I dati di input di qualità dell'aria sono rappresentati nella tabella che segue.

Variabile meteo	Tipo di variabile	Risoluzione spaziale	Risoluzione temporale	Serie storica minima	Fonte dati
$PM_{2.5}$	numerica	puntuale	1 h	17.02.2020 31.05.2020	Washington Air Quality Advisory system
BC (Black carbon)	numerica	puntuale	1 h	vedi sopra	vedi sopra
NO	numerica	puntuale	1 h	vedi sopra	vedi sopra
NO_X	numerica	puntuale	1 h	vedi sopra	vedi sopra
NO_2	numerica	puntuale	1 h	vedi sopra	vedi sopra
CO	numerica	puntuale	1 h	vedi sopra	vedi sopra
UFPs (particolato ultrafine)	numerica	puntuale	1 h	vedi sopra	Washington UNI

Alcune note sui dati:

- Il formato dei dati non è disponibile.
- I dati di $PM_{2.5}$, BC, NO, NO_X, NO_2, CO provengono da una Roadside air monitoring station (Washington Air Quality Advisory system)
- I dati di UFPs: provengono da University of Washington
- Si applicano criteri di esclusione dati sulla base di:
    - Associati ad errori.
    - Concentrazione con valore negativo.
    - Serie temporali di misure orarie incomplete (di numero dati validi inferiori al 50%).

iii. <u>Altri dati</u>

Variabile meteo	Tipo di variabile	Risoluzione temporale	Serie storica minima	Fonte dati
Hourly total vehicle volume	numerica	1 h	17.02.2020 31.05.2020	WSDOT Washington state dept. Transportation
Road occupancy: traffic volume divided by speed	numerica	1 h	vedi sopra	vedi sopra

Alcune note sui dati:

- Il formato dei dati non è disponibile.
- Si applicano criteri di esclusioni dati sulla base di:
    - Associati ad errori.
    - Volume or occupancy con valori negativi.
- Misure lockdown prese sono le Washington Stay-Home Order (SHO).

b. Metodo statistico normalizzazione

Comparazione empirica dei dati pre-post SHO (Stay-Home-Order) period (Sect 2.2)

- Dati raggruppati in 15 settimane. La settimana 0 corrisponde alla settimana prima di SHO.
- I dati delle 15 settimane, traffico, inquinamento, meteo, sono stati raggruppati in base al parametro "datetime".
- Nessuna delle variabili è risultata distribuita normalmente (Shapiro-Wilk tests). Perciò è stato condotto il test di **Wilcoxon test** su ciascuna distribuzione per i periodi pre-post SHO**.**
- R cran PACKAGES: **coin, rstatix.**

Predizione del modello sull'impatto del COVID-19 su gli inquinanti dipendenti dal traffico TRAF (sec 2.3)

L'autocorrelazione parziale è stata calcolata per ogni inquinante (PACF), da cui l'ordine di P ==1. Di seguito si indica con MAR(1) il modello autoregressivo multivariato di ordine pari a 1 (ossia si usa l' osservazione precedente più vicina).

La concentrazione degli inquinanti è stata trasformata in variabile logaritmica.

L'equazione del modello applicato a ciascun inquinante:

$$log(y_t) = \beta_0 + \beta_1 log(y_{t-1}) + \beta_2 traffic_t + \beta_3 T_t + \beta_4 RH_t + \beta_5 P_t + \beta_6 WS_t + \beta_7 WD_t + \varepsilon$$

β_0 .. β_7 sono i coefficiente di MAR(1), ε rappresenta i residui. In particolare:

- È stato scelto il parametro Road occupacy come indicatore maggiore per il traffico.
- WD può essere pari a 0, quando è prevalentemente dalla direzione ovest; mentre è pari a 1 quando risulta prevalentemente dalla direzione est.
- Gli outliers sono evidenziati tramite il valore della distanza di COOK's (esclusione per $d_{cook} > 0.5$)

Per determinare il cambiamento percentuale della mediana dell'inquinante dovuto al periodo di COVID-19 si è utilizzato il coeff $\beta_2$stimato con il modello MAR(1):

$$\Delta y(\%) = \left(e^{\beta_2 \times \Delta_{traffic}} - 1\right) \times 100\%$$

Dove $\Delta_{traffic}$ è il valore assoluto del cambiamento nella mediana del parametro occupazione della strada, dovuta alla risposta COVID-19. $\Delta_{traffic}$ è stato calcolato come la differenza tra la mediana dei valori orari di traffico per le settimane post- SHO e la settimana -2.

$$\Delta_{traffic} = traffic_{post-SHO} - traffic_{week(-2)}$$

$traffic_{post-SHO}$: mediana dei dati orari di occupazione strada per le settimane 1-10
$traffic_{week(-2)}$: mediana dei dati orari di occupazione strada per la settimana -2

Sensitivity analysis:

- Calcolo della rosa dei venti per la valutazione dei contributi da emissioni regionali.
- Sono stati considerati cinque modelli derivanti dal primo in cui di volta in volta si sono esclusi dei parametri.

- I pacchetti R utilizzati sono: **dat.table, stats, mgcv, coin, rstatix, tidyverse, openair, ggpubr, leaflet.**

i. Validazione del modello

La validazione è effettuata su:

- Calcolo della significatività dei coefficienti beta (p value).
- Calcolo del coefficiente di determinazione (R^2).
- Calcolo dell'autocorrelazione nei residui per ogni inquinante.
- Utilizzo di Aikake, Bayes information criteria.

c. Analisi criticità e punti di forza

Punti di forza:

- gli effetti di correlazione nelle serie temporali osservate giocano un ruolo importante nel modello MAR(1) spiegando il 50-80% dei livelli misurati.
- I risultati sono ottenuti basandosi sulle mediane Quindi dovrebbero risultare più robusti.
- Il modello MAR(1) fondato su dati orari può quasi eliminare il bias causato dall'assenza di dati UFP_S mancanti, poiché il range diurno di variazione del traffico ha portato un ampio intervallo agli input del modello.

Criticità:

- non vi è inclusione di altre sorgenti di inquinamento oltre il traffico.
- Il modello MAR(1) tende a sottostimare l'impatto del valore corrente del traffico sul valore dell'ora successiva.
- I risultati ottenuti sono estremamente locali (sito-specifici).

Risultato chiave: tenendo conto della meteorologia si sono ottenute delle risposte associate al periodo di COVID-19 in cui vi sono state maggiori riduzioni in UFPs relative al traffico rispetto al $PM_{2.5}$ nella regione di Seattle (WA). Questo è in contrasto con quanto rilevato dal confronto empirico diretto dei dati.

Si consideri di leggere anche:

Bekbulat, B., Apte, J.S., Millet, D.B., Robinson, A., Wells, K.C., Marshall, J.D., 2020. $PM_{2.5}$ and ozone air pollution levels have not dropped consistently across the US following so- cietal COVID response. ChemRxiv. Cambridge: Cambridge Open Engage; 2020; this content is a preprint and has not been peer-reviewed.

Chatfield C., The Analysis of Time Series. Chapman and Hall, 1996.

Penny W. and Harrison L., 2006, Chapter 40: Multivariate autoregressive models https://www.fil.ion.ucl.ac.uk/~wpenny/mbi/mar.pdf (visitato il 13 ottobre 2020).

2.2 Scheda analisi bibliografia Autoregr. n°2

Cameletti M. (2020). The Effect of Corona Virus Lockdown on Air Pollution: Evidence from the City of Brescia in Lombardia Region (Italy). *Atmos. Environ.*, *239*, 117794, ISSN 1352-2310. https://doi.org/10.1016/j.atmosenv.2020.117794.

Metodo statistico per la normalizzazione e stima del contributo

a. Dati di input

i. Dati meteo

I dati di input meteorologici sono rappresentati nella tabella che segue.

Variabile meteo	Tipo di variabile	Risoluzione spaziale	Risoluzione temporale	Serie storica minima	Fonte dati
W-speed	numerica	puntuale	< 1 h	01.01.2020 27.03.2020	ARPA LOMBARDIA Pastori-BS
Temperature	numerica	puntuale	< 1 h	vedi sopra	Pastori-BS Via Ziziola-BS
Precipitazione cumulata	numerica	puntuale	1 d	vedi sopra	Pastori-BS

Alcune note sui dati:

- Il formato dei dati non è disponibile.
- Il periodo temporale è suddiviso in due intervalli:
  - 01.01.2020 -> 07.03.2020
  - 08.03.2020 -> 27.03.2020
- È effettuato il calcolo della correlazione tra le serie temporali (Pearson).

ii. Dati qualità dell'aria delle stazioni di monitoraggio

I dati di input di qualità dell'aria sono rappresentati nella tabella che segue.

Variabile	Tipo di	Risoluzion	Risoluzione	Serie storica	Fonte dati

AQ	variabile	e spaziale	temporale	minima	
PM_{10}	numerica	puntuale	1 h	01.01.2020 27.03.2020	ARPA LOMBARDIA 1. Broletto 2. Villaggio Sereno
NO_2	numerica	puntuale	1 h	vedi sopra	1. Broletto 2. Via Turati 3. Villaggio Sereno

iii. Altri dati

variabile	Tipo di variabile	Risoluzione temporale	Serie storica minima	Ulteriori caratteristiche
Sunday effect	cat	1 d	17.02.2020 31.05.2020	
Stazione Broletto	num	1 h/1d	Vedi sopra	Traffic Urbano
Stazione Via Turati	num	1 h/1d	Vedi sopra	Traffic Urbano
Stazione Villaggio Sereno	num	1 h/1d	Vedi sopra	BG urbano

b. Metodo statistico normalizzazione

Modello utilizzato (sez. Introduzione e Metodologia)

Viene utilizzato un modello di serie temporali interrotte (ITS, o regressione segmentata) (Wagner et al., 2002). Questo tipo di approccio modellistico è ampiamente utilizzato per la valutazione dell'impatto longitudinale degli interventi pubblici o di importanti azioni relative alla popolazione (Grundy et al., 2009; Bernal et al., 2009).

Nel caso di studio considerato, la variazione del livello e/o l'andamento delle concentrazioni di PM_{10} e NO_2 viene modellata includendo in un modello di regressione lineare una variabile fittizia e un termine di interazione. Inoltre, per regolare la stagionalità e l'autocorrelazione nella variabile di risposta, le variabili meteorologiche sono incluse nel modello e si assume una struttura autoregressiva a media mobile (ARMA) per il termine di errore. Nella letteratura sulle serie temporali questo tipo di modello è noto anche come

modello ARMAX (Hyndman e Athanasopoulos, 2019). L'analisi viene implementata utilizzando il pacchetto di **forecast** del software R (R Core Team, 2020).

Una regressione di serie temporali interrotta è stata impiegata per la comparazione del periodo precedente (1° gennaio-7 marzo 2020) e successivo (8 marzo-27 marzo 2020) la misura di contenimento.

Sia y_t la concentrazione di inquinante (PM_{10} o NO_2) al tempo t (t = 1, ..., n) misurata da una singola stazione di monitoraggio. Inoltre, L_t rappresenta la variabile fittizia lockdown (L) che è uguale a 0 per tutti i punti temporali prima del blocco e 1 dopo (ovvero dall'8 marzo 2020 al 27 marzo 2020). La variabile T = 1, ..., n rappresenta il tempo trascorso (in giorni) a partire dal 1° gennaio 2020 che è il primo punto temporale considerato nell'analisi (n = 87 è il numero totale di giorni). Il vettore X_t di dimensioni 1 × p contiene i valori dei regressori considerati riferiti al giorno t dati da temperatura, precipitazioni, velocità del vento e una variabile fittizia introdotta per catturare l'effetto domenica (la variabile è uguale a 1 se il giorno t è avvenuto di domenica, e 0 altrimenti). La scelta di includere una variabile per la domenica, invece di una per l'effetto weekend (sabato – domenica), si basa su un'ispezione visiva della stagionalità delle serie temporali.

Il modello applicato a tutte e cinque le serie temporali di inquinanti è il seguente:

$$y_t = \alpha_0 + \alpha_1 T + \alpha_2 1_t^L + \alpha_3 T 1_t^L + X_t\beta + \eta_t$$
$$\eta_t = \phi_1\eta_{t-1} + \cdots + \phi_p\eta_{t-p} + \epsilon_t + \theta_1\epsilon_{t-1} \ldots + \theta_q\epsilon_{t-q}$$

- Prima equazione contiene: intercetta, effetto lineare del tempo, regressori ed effetto lockdown
- Seconda equazione contiene: struttura temporale degli errori di regressione.

In particolare, il termine di errore di regressione η_t segue un processo ARMA con coefficienti ϕ_1, ..., ϕ_p, θ_1, ..., θ_q, mentre il termine di errore di innovazione ϵ_t è assunto come un processo di rumore bianco normalmente distribuito a media zero e varianza σ^2.

In questo modello:

- $\boldsymbol{\beta}$ è il vettore $p \times 1$ dei coefficienti per i regressori.
- Il termine $\alpha_0 + X_t\boldsymbol{\beta}$ rappresenta il livello di base della variabile di risposta quando $1L_t = 0$ (prima del blocco)
- α_1 è la pendenza del trend di base, cioè la variazione attesa di y_t che si verifica ogni giorno prima dell'intervento.

Quando il lockdown ha effetto, il livello e la pendenza del trend diventano uguali rispettivamente a $\alpha_0 + \alpha_2 + X_t\boldsymbol{\beta}$ e $\alpha_1 + \alpha_3$. Pertanto, $\alpha 2$ ($\alpha 3$) rappresenta la variazione del livello post-intervento (pendenza del trend).

Il vettore finale dei parametri è dato da ($\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, ϕ_1, ..., ϕ_p, θ_1, ..., θ_q, σ_2), con $\boldsymbol{\alpha}$ = (α_0, α_1, α_2, α_3) e l'approccio di massima verosimiglianza viene utilizzato per la stima (Hyndman e Athanasopoulos, 2019).

In particolare si adotta un metodo graduale all'indietro partendo dal modello completo e rimuovendo i regressori non significativi (tra temperatura, precipitazione, velocità del vento e variabile dummy Sunday) in base al valore del valore p (la soglia considerata è α = 0,05). I coefficienti ARMA vengono mantenuti nel modello anche se non significativi perché migliorano il fit del modello. Anche i parametri $\boldsymbol{\alpha}$ non vengono rimossi dal modello per discutere l'efficacia dell'intervento di lockdown nel modificare il livello e la pendenza del trend delle serie temporali.

Se le informazioni di una qualche variabile sono mancanti, si assume che i coefficienti non si discostino molto da zero e il termine si rimuove dal modello.

i. <u>Validazione del modello</u>

Si utilizza il Akaike Information Criterion (AIC) per selezionare Il migliore ARMA model in termini di determinazione/minimizzazione dell'errore. Inoltre, si analizzano i residui per assicurarsi che gli errori ARMA ϵt assomiglino a una serie temporale di rumore bianco.

Risulta in aggiunta (sez. risultati):

- La temperatura non ha un'influenza significativa sulle concentrazioni ed è stata rimossa.
- La velocità del vento è la variabile più rilevante sulle concentrazioni (effetto di riduzione all'aumentare dell'intensità).
- La variabile dummy *Sunday* risulta sempre diversa da zero ad eccezione del caso di un sensore di PM_{10} collocato nel centro di Brescia.
- L'effetto maggiore di lockdown si è misurato per un sensore di NO_2 collocato in una stazione di traffico urbano.
- Gli errori ARMA risultano distribuiti in modo non significativamente differente alla distribuzione normale del rumore bianco.

c. Analisi criticità e punti di forza

Punti di forza:

- Vengono rappresentate le serie temporali del periodo di lockdown insieme alle previsioni controfattuali (no lockdown e assunzione di serie temporale stazionaria), per cui è possibile paragonare due scenari diversi.

Criticità:

- La precipitazione ha un effetto significativo solo sulle misure di un sensore di NO_2. È molto strano però che non ci siano effetti sul PM_{10}.

Si consideri di leggere anche:

Bernal, J.L., Cummins, S., Gasparrini, A., 2016. Interrupted time series regression for the evaluation of public health interventions: a tutorial. Int. J. Epidemiol. 46 (1), 348–355.

Grundy, C., Steinbach, R., Edwards, P., Green, J., Armstrong, B., Wilkinson, P., 2009. Effect of 20 mph traffic speed zones on road injuries in London, 1986-2006: controlled interrupted time series analysis. BMJ 339.

Hyndman, R., Athanasopoulos, G., 2019. Forecasting: Principles and Practice, third ed. OTexts, Melbourne, Australia, URL https://otexts.com/fpp3/. (Accessed 30 March 2020).

R Core Team, 2020. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, URL https://www.R-project. org/.

Wagner, A.K., Soumerai, S.B., Zhang, F., Ross-Degnan, D., 2002. Segmented regression analysis of interrupted time series studies in medication use research. J. Clin. Pharmacy Therapeut. 27 4, 299–309.

Cameletti, M., Ignaccolo, R., Bande, S., 2011. Comparing spatio-temporal models for particulate matter in piemonte. Environmetrics 22 (8), 985–996.

Maranzano, P., Fassò, A., Pelagatti, M., Mudelsee, M., 2020. Statistical modeling of the early-stage impact of a new traffic policy in Milan, Italy. Int. J. Environ. Res. Publ. Health 17, 1088.

2.3 Sintesi modelli di autoregressione multivariata

I modelli autoregressivi multivariati si basano sul concetto che le serie temporali di variabili osservate, costituite da misurazioni consecutive delle grandezze in esame, contengono informazioni sui processi che le hanno generate. In questo approccio si ipotizza che il valore corrente della variabile sia rappresentato come somma lineare ponderata dei suoi valori precedenti, in quanto si suppone che vi sia una correlazione temporale tra i valori a istanti diversi e anche di altre grandezze. In questo contesto l'ordine del modello è il numero di osservazioni utilizzate precedenti nel tempo e i pesi caratterizzano le serie temporali fornendo in aggiunta una stima della significatività dei diversi regressori.

La formulazione dei modelli di autoregressione multivariata è in genere espressa come segue.

Si consideri un numero -*d*- di serie temporali generate da *d*-variabili all'interno di un sistema dove *m* è l'ordine del modello. Un modello MAR (m) prevede il valore successivo in una serie temporale *d*-dimensionale, y_n, come combinazione lineare dei *m* valori vettoriali precedenti:

$$y_n = \sum_{i=1}^{m} y_{n-1} A(i) + e_n$$

dove $y_n = [y_n(1), y_n(2), ..., y_n(d)]$ è l'ennesimo campione di una serie temporale d-dimensionale, ogni A (i) è una matrice di coefficienti (pesi) ed $e_n = [e_n(1), e_n(2), ..., e_n(d)]$ è il rumore gaussiano additivo con media zero e covarianza R. Si noti che si è assunto che la media dei dati sia stata sottratta dal tempo serie.

I modelli autoregressivi multivariati vengono utilizzati nello studio dell'inquinamento atmosferico per associarlo a cause specifiche come l'influenza delle variabili meteorologiche, del tempo, delle emissioni di inquinanti in atmosfera. Ad esempio, nello studio di Xiang et al. (2020), un modello autoregressivo multivariato è stato applicato per correggere eventuali correlazioni e per studiare l'influenza delle variabili meteorologiche (temperatura, umidità relativa, precipitazione, velocità e direzione del vento) e del traffico sulle concentrazioni d'inquinanti (BC, $PM_{2.5}$, NO_X, CO, UFPs) misurate in periodi prima e dopo l'applicazione di misure di distanziamento per l'emergenza COVID-19 negli Stati Uniti (17 febbraio-31 marzo 2020).

I modelli autoregressivi multivariati possono essere utilizzati anche in associazione ad altri modelli. Come esempio, lo studio di Cameletti 2020 ha lo scopo di valutare l'efficacia delle misure di distanziamento introdotte dal governo italiano per contrastare l'emergenza COVID-19 sul miglioramento della qualità dell'aria (in particolare NO_2 e PM_{10}). In particolare, viene presentato un metodo che è basato sull'analisi della regressione segmentale per testare un eventuale cambio significativo nel livello e nell'andamento degli inquinanti dovuto alle misure di Lockdown. Il modello proposto si avvale della segmentazione delle serie temporali di dati di concentrazione di inquinanti su due periodi (1-gennaio al 7 marzo 2020; 8 marzo 2020- 27 marzo 2020) che rappresentano l'assenza e la presenza delle misure di contenimento. In particolare, il modello è qui rappresentato da due equazioni. La prima rappresenta un'equazione di regressione, che include l'effetto lineare del tempo, variabili metereologiche

(temperatura, vento, precipitazione) e categoriche. La seconda equazione definisce invece la struttura temporale degli errori di regressione e consiste in un modello autoregressivo a media mobile. Qui si assume che l'errore, che rappresenta la variabilità arbitraria non espressa dal modello, sia distribuito normalmente e sia correlato con valori precedenti ad istanti diversi.

3 MODELLI RANDOM FOREST

Il metodo di Random Forest (RF) (foresta delle decisioni) è una tecnica di Machine learning ad alberi decisionale di insieme. Gli alberi decisionali utilizzano un algoritmo binario di classificazione ricorsiva che crea nodi "puri" tramite la divisione delle osservazioni in due gruppi omologhi. La natura ricorsiva dell'algoritmo significa che la divisione viene ripetuta fino a quando non viene raggiunta la purezza del nodo. Gli alberi decisionali sono inclini all'overfitting pertanto, per ovviare a questo svantaggio vengono fatti “crescere” tanti alberi decisionali individuali da un set di addestramento che utilizza un processo chiamato bagging (aggregazione bootstrap). (Grange et al. 2018. § 1.3. pag. 4/18).

Il bagging si riferisce alla sostituzione casuale delle osservazioni campionate dal training set insieme al campionamento di variabili esplicative. I dati ottenuti da ciascuna operazione di bagging sono chiamati dati out-of-bag (OOB). Quando viene coltivato un singolo albero dai dati OOB, se il processo viene ripetuto, è improbabile che contengano le stesse osservazioni e variabili utilizzate da altri alberi. I modelli RF di solito contengono alcune centinaia di alberi utilizzando i dati OOB e questo crea una foresta che consiste di molti alberi non correlati che sono stati addestrati su diversi sottoinsiemi di training set. Ogni albero può quindi essere utilizzato per fare previsioni, che vengono aggregate per formare una singola previsione. Nelle applicazioni di regressione, viene utilizzata la media delle previsioni. Ciò consente alla RF di produrre modelli predittivi che generalizzano bene, e la prestazione predittiva è generalmente considerata tra le migliori di qualsiasi tecnica ML.

3.1 Scheda analisi bibliografia RF n°1

Grange, S. K., & Carslaw, D. C. (2018). Using meteorological normalisation to detect interventions in air quality time series Science of the Total Environment Using meteorological normalisation to detect interventions in air quality time series. *Science of the Total Environment*, *653*(November), 578–588. https://doi.org/10.1016/j.scitotenv.2018.10.344

Grange, S. K., Carslaw, D. C., Lewis, A. C., Boleti, E., & Hueglin, C. (2018). Random forest meteorological normalisation models for Swiss PM_{10} trend analysis. *Atmospheric Chemistry and Physics*, 18, 6223–6239. https://doi.org/10.5194/acp-18-6223-2018

Metodo statistico per la normalizzazione e stima del contributo

a. Dati di input

i. Dati meteo

Variabili meteo	Risoluzione spaziale	Risoluzione temporale	Serie storica min.	Fonti dati	Aggregazione minima	Formato dati
Wind Speed*	Puntale	h	(1997–2016)	NOAA's (ISD)	d	.csv
Wind Direction	Puntale	h	(1997–2016)	NOAA's (ISD)	d	.csv
Atmospheric Temperature	Puntale	h	(1997–2016)	NOAA's (ISD)	d	.csv
Humidity	Puntale	h	(1997–2016)	NOAA's (ISD)	d	.csv
Pressure	Puntuale	h	(1997–2016)	NOAA's (ISD)	d	.csv
PBL	0.125x 0.125(°)	d	(1997–2016)	(ECMWF) Portale dati ERA-Interim	d	
Back trajectory	Puntale	d	(1997–2016)	HYSPLIT model	5 d	
scale weather patterns (WTC)	Regioni Alpine	d	(1957–2011)	ECMWF ERA40 ERA IFS	stagionale	

Risoluzione Spaziale: h(oraria), d (giornaliera).
**velocità scalare*

ii. Dati qualità dell'aria

Inquinanti	Risoluzione temporale	Serie storica min.	Fonti dati	Aggregazione minima	Tipo Stazione	Formato dati
PM_{10}	d	(1997–2016)	(EEA) Air-Base (AQER) Air Quality e-Re-porting (FOEN) Swiss Federal Office for the Envi-ronment	d	R, RMW*, RMT*, UT, UB, SU	.csv
NOx	h	(1997–2016)	smonitor Europe	d	UT	.csv
NO_2	h	(1997–2016)	smonitor Europe,	d	UT	.csv
SO_2	h	(2001–2012)	Kent Air Quality data-base	d	UB, UI	.csv

Dati Qualità dell'aria. Classificazione non tipo EOI.

**RMW (Rural Motor Way), *RMT (Rural Mountain).*

I dati delle concentrazioni degli inquinanti considerati sono stati modellati solo se erano disponibili dati validi sulla velocità del vento per quel giorno. Per tutte le altre variabili di input, sono stati assegnati ai dati mancanti i valori della mediana per le variabili numeriche (Grange et al. 2018. § 2.2. pag. 6/18).

iii. Altri dati e metainformazioni.

N.ro di Variabili	Variabili
1	Julian Day
2	Week day
3	Hours

Tabella variabili esplicative e metainformazioni.

Numero di Metadati	Metadati
1	Tipo di Stazione
2	Latitudine
3	Longitudine
4	Altezza Stazione (slm)

Tabella delle metainformazioni.

iv. Analisi criticità e punti di forza.

Si evidenziano:

- Possibili criticità nell'utilizzo di serie di dati di variabili meteo con la variabile "velocità del vento "poco popolata. Con valori di WS non disponibili o uguale a zero, il metodo RF utilizzato non considera tutti i dati corrispondenti.
- Tra i punti di forza annoveriamo la possibilità di utilizzo di repository e/o dB con dati meteo e della qualità dell'aria già validati.

b. Metodo statistico normalizzazione

i. Descrizione algoritmo

In pratica per ogni sito è stato utilizzato il modello RF per predire ogni concentrazione di PM_{10} mille volte. Per ogni previsione, le variabili esplicative, ad eccezione del termine di tendenza (trend term), sono state campionate senza sostituzione e in modo casuale assegnate a un'osservazione di variabile dipendente (una concentrazione di PM_{10}). Le 1000 previsioni sono state quindi aggregate utilizzando la media aritmetica, e questa rappresenta la condizione meteorologica "media", quindi la tendenza normalizzata meteorologicamente. Adoperando più di mille previsioni è stata ottenuta solo una minima riduzione del rumore.

La formazione dei modelli è stata condotta sull'80% dei dati di input e l'altro 20% è stato trattenuto dalla formazione e utilizzato per la fase di validazione.

Paper	N° di alberi	N° di variabili per nodo	Min Dimensione nodo	Random forest model R^2
Grange et al. 2018 (§ 3.1. pag. 7/18)	300	3	5	0,54 ÷ 0,71 (PM_{10})
Grange et al. 2019 (§ 2.1.2. pag. 6/12)	300	3	5	0,67 (SO_2) 0,82 (NO_2) 0,83 (NO_x)

Formazione e valutazione del modello.

Per i dati di SO_2 della stazione di Dover, i modelli sono stati calcolati utilizzando il set osservazionale completo, ma dopo aver studiato i modelli (discussa nella sezione § 3.1.1, pag. 6/12), le osservazioni sono state filtrate in funzione della direzione del vento. Tali modelli sono stati utilizzati per l'analisi delle serie temporali (§ 3.1.1, pag. 6/12).

I pacchetti R utilizzati sono i seguenti, con la disponibilità del codice sorgente e la presenza della descrizione delle procedure:

1. **rmweather** R package (version 0.1.2) (R Core Team, 2018; Grange,2018). Grange, S.K., 2018. rmweather: Tools to Conduct Meteorological Normalisation on Air Quality Data. R package version 0.1.2. https://CRAN.R-project.org/package=rmweather.
2. **smonitor:** A framework and a collection of functions to allow for maintenance of air quality monitoring data. 2018. url: https://github.com/skgrange/smonitor.

ii. Validazione del modello (indicatori di performance utilizzati e commento dei risultati)

Il confronto dei modelli di normalizzazione meteorologica RF con altre tecniche non erano un obiettivo primario di questo lavoro. Tuttavia è importante considerare quale effetto ha avuto la normalizzazione meteorologica sulle stime di tendenza. Per indagare su questo, le osservazioni di PM_{10} sottoposte al processo di normalizzazione meteorologica sono state aggregate come medie mensili e le loro tendenze testate con il test Theil-Sen. A tale

proposto si evidenzia come questa potrebbe essere considerata una procedura "standard" per l'analisi dei dati sulla qualità dell'aria.

Per i valori di PM_{10} normalizzati in Svizzera dal 1997 al 2016, ad eccezione dei siti autostradali rurali, la stima tendenziale normalizzata è risulta essere maggiore rispetto alle stime di tendenza non normalizzate. Ciò indica che la meteorologia in Svizzera tra Il 1997 e il 2016 ha mascherato i cambiamenti nelle emissioni di PM_{10} durante lo stesso periodo. Le stime di tendenza normalizzate avevano un intervallo di incertezza molto più basso in tutti i casi analizzati.

Nel caso dell'applicazione ai dati di NO_2 e NO_x nella stazione di Londra e ai dati di SO_2 del porto di Dover, la tecnica della normalizzazione meteorologica ha permesso di investigare gli effetti degli interventi messi in atto per ridurre i livelli in aria di tali inquinanti nelle due aree, identificando gli eventi di breakpoints (utilizzato strucchange R package).

L'algoritmo della RF non offre direttamente la possibilità di determinare l'errore o l'incertezza delle stime. Per consentire di valutare l'incertezza per questi ultimi casi di studio sono stati fatti crescere 50 modelli RF per ogni esempio con set di input campionato in modo casuale (bootstrap). Il processo bootstrap ha garantito che i modelli fossero cresciuti su diversi set di training. I valori di importanza, una misura della forza delle variabili o dell'influenza sulla previsione, le dipendenze parziali e le previsioni per ciascuno dei 50 modelli sono state quindi stimate.

3.2 Scheda analisi bibliografia RF n°2

Petetin, H., Bowdalo, D., Soret, A., Guevara, M., Jorba, O., Serradell, K., Garcìa-Pando, C.P. (2020). Meteorology-normalized impact of COVID-19 lockdown upon NO_2 pollution in Spain. *Atmospheric Chemistry and Physics*, *20*, 11119–11141. https://doi.org/10.5194/acp-20-11119-2020

Metodo statistico per la normalizzazione e stima del contributo

a. Dati di input

i. Dati meteo (§ 2.2, pag.4)

Variabile meteo	Risoluzione spaziale	Risoluzione temporale	serie storica	fonte dati	formato
daily mean 2-m temperature	31 km	giornaliera	01/01/2013-13/04/2020	reanalisi ERA5 (Copernicus Climate Change Service – C3S	nd*
minimum 2-m temperature					
maximum 2-m temperature					
surface wind speed					
normalized 10-m zonal and meridian wind speed components					
surface pressure					
total cloud cover					
boundary layer height					

*nd = non disponibili

- Dati meteorologici estratti dal database delle reanalisi ERA5 (Copernicus Climate Change Service – C3S) con risoluzione spaziale di 31 km.
- Per ogni stazione di QA sono stati estratti i seguenti parametri meteo su scala giornaliera:
- Temperatura media giornaliera a 2 m;
- Temperatura massima e minima a 2 m;
- Velocità del vento superficiale;
- Componenti zonali e meridiane della velocità del vento normalizzate a 10 m;
- Pressione superficiale;
- Copertura nuvolosa;

- Altezza del boundary layer.

ii. <u>Dati qualità dell'aria</u> (§ 2.1, pagg.3-4)

Inquinante	Risoluzione temporale	serie storica	fonte dati	aggregazione dati	classificazione stazione	formato
NO_2	oraria	01/01/2013-13/04/2020	European Environmental Agency (EEA) AQ e-Reporting + database GHOST	551 stazioni	urban/suburban background + traffic	nd*

*nd = non disponibile

- Sono state considerate le misure orarie di NO_2 della rete di monitoraggio spagnola comunicati all'European Environmental Agency (EEA) Ad e-Reporting. Dati validati, E1a per il periodo 2013-2019 e dati E2a per il periodo 2020. I dati sono stati verificati anche nel progetto GHOST (Globally Harmonised Observational Surface Treatment) un progetto per l'armonizzazione di dati osservati e metadati.
- Dati orari per il periodo 01/01/2013-23/04/2020.
- Metodo di misura a chemiluminescenza
- Disponibili 551 stazioni di NO_2 per il periodo 2013-2020.
- Messo a punto un algoritmo per la selezione automatica dei dati in modo che fosse selezionata almeno una stazione urbana/suburbana di background e una di traffico per ogni NUTS-3.
- Inserite soglie diverse per la selezione della minima disponibilità di dati validi per stazione:
    - 50% di dati giornalieri per l'intero periodo di studio;
    - 50% di dati per il periodo 01/01/2017-31/12/2019 (utilizzati per il training del modello ML – machine learning);
    - 25% di dati nel periodo 01/01/2020-13/03/2020 (utilizzati per testare il modello ML);

- 10% di dati durante il periodo di lockdown.

- Le stazioni identificate che rispondevano ai criteri di selezione sono state identificate in 50 province spagnole (38 province con stazioni di background urbano e 37 con stazioni di traffico).

iii. Altri dati**:**

- Le stazioni di QA sono state selezionate per massimizzare sia la densità di popolazione circostante (con raggio geodetico di 5 km) che la disponibilità di dati.
- Densità di popolazione relativa ad ogni stazione di QA ricavata dal database GHOST (Gridded Population of the World versione 5 – GPW – sono dati pubblici e disponibili).
- Tra le variabili esplicitato anche date index, Julian date e weekday.

iv. Analisi **criticità e punti di forza** (dettagli in Appendice A, pag. 20)

+ Effettuata analisi di Quality Assurance al dataset di dati di NO_2 utilizzando il database GHOST (vedi Appendice A del manoscritto, pag. 20). In particolare sono stati rimossi dati mancanti; dati infinito; misure negative; misure nulle; misure associate con alti livelli di incertezza e bias (stabilita nel database GHOST); misure per cui non rimanessero sufficienti dati validi da mediare; misure con persistenti valori ricorsivi; misure con valori maggiori del limite possibile (superiore a 5000 ppbv); misure che mostrano outliers da analisi con boxplot; misure identificate con valori troppo estremi; misure al di sotto del limite inferiore di rilevamento.

b. Metodo statistico normalizzazione

i. Descrizione algoritmo (para 2.3, pagg.5-7 + Appendice C):

- Utilizzato il metodo del Gradiente Boosting Machine (GBM), decision tree-based ensemble appartenente al boosting family.

- I dati storici più recenti utilizzati per riprodurre il mixing ratios di NO_2 utilizzando temperatura media giornaliera a 2 m; temperatura massima e minima a 2 m; Velocità del vento superficiale; Componenti zonali e meridiane della velocità del vento normalizzate a 10 m; Pressione superficiale; Copertura nuvolosa; Altezza del boundary layer; date index (giorni dal 01/01/2013); Julian day e weekday. Dati giornalieri.
- GBM addestrato negli ultimi 3 anni completi (2017-2019) e poi utilizzato per predire i mixing ratios business-as-usual di NO_2 per i seguenti 4 mesi del 2020.
- Il metodo ML è stato fatto girare con script in Python usato il pacchetto *scikit-learn*.
- Il learning rate fissato a 0.05. Nel pacchetto in Python selezionata opzione "sqtr" per il max_features.
- Condotta analisi sull'ottimo hyperparameter tuning: *max_depth (values from 1 to 5 by 1; subsample from 0.3 to 1.0 by 0.1; n_estimators from 50 to 1000 by 50; min_samples_leaf from 1 to 30)*.
- Sui principali iperparametri condotta una "randomized search". Esplorate 20 possibili combinazioni di hyperparameters.

ii. Validazione del modello (§ 2.3.3, pagg. 6-7 + para 3.1, pagg.7-9)

- Non sono stati notati miglioramenti se il periodo di addestramento è ampliato a 4 o 5 anni.
- Utilizzati i primi due mesi e mezzo del 2020 per validare il modello.
- Condotta stima incertezza replicando lo stesso esperimento addestrando il ML in 4 differenti periodi storici: 2013-2015; 2014-2016; 2015-2017 e 2016-2018 e testando i risultanti nei primi 4 mesi del 2016, 2017, 2018 e 2019, rispettivamente.
- Ottenuti per ogni stazione 538 residui giornalieri, da cui il 5° e 95° percentile rappresentano l'intervallo di incertezza. Per ogni stazione,

ottenuto un intervallo di confidenza asimmetrico fissato al 90% utilizzato per caratterizzare l'incertezza nella predizione dei primi 4 mesi del 2020,

- Mediando su tutte le province spagnole, l'intervallo di incertezza varia da [-5.5,+5.2] ppbv alle stazioni di fondo urbano e [-6.7,+6.6] ppbv nelle stazioni di traffico.
- Oltre analisi incertezza giornaliera, anche analisi incertezza settimanale. Stesso metodo precedente ma con i seguenti risultati: [-4.2,+3.5] ppbv alle stazioni di fondo urbano e [-4.9,+4.5] ppbv nelle stazioni di traffico, che rappresentano rispettivamente un riduzione del 27% e 30% rispetto all'incertezza giornaliera.
- Per la validazione dei risultati sono state utilizzate le seguenti metriche: mean bias (MB), normalized mean bias (nMB), root mean square error (RMSE), normalized root mean square error (nRMSE) e Person correlation coefficient (PCC).
- Per il periodo di training del modello (01/01/2017-31/12/2019) nelle stazioni di fondo urbano non sono stati osservati bias, valori di RMSE sotto il 30% (20% come media sulle province) e PCC in media pari a circa 0.91. Risultati simili anche per le stazioni di traffico.
- Test sul periodo precedente il lockdown (01/01/2020-13/03/2020), le prestazioni sono buone in molte province: bias incrementa al +2% (FU) e +7% (TRA), RMSE al 32% (FU) e 28% (TRA), PCC si riduce a 0.71 (FU) e 0.75 (TRA). Notata variabilità interregionale tra le province.
- **RISULTATI** (para 3.3, pagg. 13-14): le concentrazioni di NO_2 sono diminuite durante il periodo di lockdown per entrambe le tipologie di stazione, in particolare per FU riduzione di circa -4[-8,+0] ppbv (in termini relativi pari a circa -50[-94,+3]%; per TRA riduzione di circa -6[-11,-1] ppbv (in termini relativi -50[-88,-8]%).

- Notate differenze nelle riduzioni tra le diverse fasi del lockdown: riduzione in media pari a circa il 40% per entrambe le tipologie di stazione nella fase I del lockdown, che aumenta al 55% nelle fasi II e III.

iii. Analisi criticità e punti di forza (analisi dettaglio in § 3.5, pagg. 17-18)

+ Valutazione dell'incertezza: utilizzata incertezza ottenuta su stima settimanale che secondo gli autori potrebbe essere probabilmente conservativa nella stima dell'intero periodo di lockdown.

+ Statistiche di validazione (MB, RMSE o PCC) con risultati scarsi nelle stazioni di alcune province, dovuta ad una probabile difetto nel modello ML. Nonostante tutti le procedure di tuning del modello applicate, resta qualche sovrastima. Tra le possibili spiegazioni anche l'utilizzo per i dati meteo del database ERA5 (31 km di risoluzione) che per alcune stazioni a orografia complessa potrebbero non essere sufficientemente adeguati.

c. Metodo statistico per la stima del contributo delle misure attuate

i. Descrizione algoritmo (para 3.4, pagg. 14-17):

- La stima del contributo delle misure adottate durante il lockdown, una volta depurato l'effetto delle concentrazioni da parte della meteo, non è stato valutato con alcun algoritmo.
- Analisi qualitativa e ipotesi relazione lineare tra le concentrazioni di NO_2 registrate nelle stazioni di fondo urbano e le emissioni locali di NO_X stimate su una griglia di 4 km x 4 km.
- Utilizzato inventario delle emissioni antropogeniche bottom-up disponibile sull'intera Spagna con una risoluzione spaziale orizzontale di 4km x 4km, stimate col modello emissivo HERMESv3, disponibile al

seguente repository gitlab repository: https://earth.bsc.es/gitlab/es/hermesv3_bu

ii. <u>Validazione del modello</u>

- L'ipotesi di relazione lineare tra concentrazione NO_2 ed emissioni NO_X nelle aree considerate ha portato a stimare una riduzione emissiva di NO_X del 70%(80%) dal settore del trasporto stradale con conseguente riduzione di NO_2 del 47%(54%). Gli altri settori potenzialmente coinvolti dal lockdown (settore industriale e residenziale) non sono stati stimati in questa fase e si rimanda ad un paper successivo per approfondimento (attualmente in discussione sempre su ACP, Guevara et al., 2020: Time-resolved emission reductions for atmospheric chemistry modelling in Europe during the COVID-19 lockdowns").

3.3 Scheda analisi bibliografia RF n°3

Kaminska, J.A. (2018). The use of random forests in modelling short-term air pollution effects based on traffic and meteorological conditions: A case study in Wroclaw. *Journal of Environmental Management, 217*, 164-174.

Metodo statistico per la normalizzazione e stima del contributo

a. Dati di input

i. Dati meteo (§ 2.3, pagg. 5-6)

Variabile meteo	Risoluzione spaziale	Risoluzione temporale	serie storica	fonte dati	formato
air temperature	puntuale (una stazione meteo)	oraria	01/01/2015-31/12/2016	Institute of Meteorology and Water Mangement (IMGW)	nd*
wind speed	puntuale (una stazione meteo)	oraria			
wind direction	puntuale (una stazione meteo)	oraria			
relative humidity	puntuale (una stazione meteo)	oraria			
atmospheric pressure	puntuale (una stazione meteo)	oraria			

*nd = non disponibile

ii. Dati qualità dell'aria (§ 2.2, pagg. 3-5)

Inquinante	Risoluzione temporale	serie storica	fonte dati	aggregazione dati	classificazione stazione	formato
NO_2	oraria	01/01/2015-31/12/2016	Provincial Environment Protection Inspectorate	3 stazioni	traffico	nd*
$PM_{2.5}$	oraria			2 stazioni	traffico	

*nd = non disponibile

iii. Altri dati

- Le variabili considerate sono state suddivise in nove periodi interessanti per esaminare i diversi comportamenti del metodo:
    - intero periodo di 2 anni;
    - combinazione delle stagioni calde (aprile/settembre) di entrambi gli anni (2015 e 2016);
    - combinazione delle stagioni fredde (ottobre/marzo) di entrambi gli anni (2015 e 2016);
    - giorni lavorativi;
    - giorni festivi;
    - primavera (marzo-maggio);
    - estate (giugno-agosto);
    - autunno (settembre-novembre);
    - inverno (dicembre-gennaio).
- Essendo stazioni da traffico, è stato considerato il volume di traffico;
- La variabile direzione del vento non è stata espressa solo in gradi (per evitare per esempio sovrapposizione di 1° e 360°) per tale motivo la direzione del vento è stata espressa usando le 8 categorie con separazione di 45° (N, NE, E, ecc.).

iv. Analisi criticità e punti di forza.

- Lunghezza della serie storica (considerata serie di due anni)

b. Metodo statistico normalizzazione

i. Descrizione algoritmo (§ 2.4, pagg. 6-8)

- Applicato metodo Random Forest (RD): l'importanza delle variabili predittive è stata determinata come somma, per ogni nodo dell'albero, degli incrementi del paramento ΔR (resubstitution estimate): valore espresso come percentuale sul massimo della

somma di tutte le variabili. ΔR è stato ottenuto come somma di tutti i predittori su tutti i nodi e gli alberi

ii. Validazione del modello (§ 2.4, pagg. 6-8)

- Il modello costruito è stato addestrato su tutti e nove i periodi considerati utilizzando come training set il 50% di tutto il campione e il 30% del campione per il test. Il processo di apprendimento (learning process) si è fermato quando per 10 cicli l'errore era inferiore al 5%.
- Il processo precedente ha determinato il numero di alberi in soli due casi: NO_X e $PM_{2.5}$ quando si è considerato l'intero periodo di due anni. Il numero di variabili individuato è pari a 9, il numero di predittori selezionati per la costruzione di un albero sono stati 5 e il conseguente numero di subsets variables pari a 126, per cui il numero di alberi è stato limitato ad un massimo di 200;
- Per valutazione correlazione utilizzato il coefficiente di determinazione R^2. Per due coppie di variabili si sono trovate dipendenze che sono state approfondite: umidità relativa e temperatura; umidità relativa e volume di traffico.
- La validazione del modello è stata effettuata considerato come coefficienti R^2, MFB (Mean Fractional Bias), MAD (Mean Absolute Deviation error) e MAPE (Mean Absolute Percentage Error), mentre criteri come BIC e AIC non sono stati considerati perché il numero di variabili nel modello sono predefinite e costanti.
- Alcuni risultati: valore di R^2 generalmente basso (sotto il valore di 0.57); MFB meno di 0.2; il valore di MAPE indica prestazioni migliori per NO_2 che $PM_{2.5}$. La suddivisione in periodo ha mostrato risultati migliori nella stazione estiva e nel periodo caldo. Il tipo di giorno (lavorativo e non) non sembra avere importanza.

iii. Analisi criticità e punti di forza

+ per ogni coppia di variabili è stata investigata l'esistenza di collinearità;

- non è stata testata la significatività statistica;

+ validazione dettagliata del modello con alcuni indicatori statistici (MFB, R^2, ecc)

c. Metodo statistico per la stima del contributo delle misure attuate

i. Descrizione algoritmo

Non applicato per la stima del contributo delle misure di distanziamento fisico. Nel caso oggetto di studio, l'importanza di ogni variabile è stata stimata in termini di *share of validity (SoV) – the percentage contribution to the total sum of importance in a given model.*

3.4 Scheda analisi bibliografia RF n°4

Mallet, M.D. Meteorological normalisation of PM_{10} using machine learning reveals distinct increases of nearby source emissions in the Australian mining town of Moranbah. *Atmospheric Pollution Research,* in press (2020).

Metodo statistico per la normalizzazione e stima del contributo

a. Dati di input

i. Dati meteo (§2.2.2, pagg. 3-4)

Variabile meteo	Risoluzione spaziale	Risoluzione temporale	serie storica	fonte dati	formato
Temperature (°C)	puntuale (Moranbah meteo station)	oraria	2011-2019	Queensland Government	nd*
Wind speed (m/s)	puntuale (Moranbah meteo station)	oraria	2011-2019		
Wind direction (°)	puntuale (Moranbah meteo station)	oraria	2011-2019		
Pressure (hPa)	31 km	oraria	2011-2019	ERA5 reanalysis	
Boundary layer height (m)	31 km	oraria	2011-2019		
Rainfall (mm/d)	puntuale (2 siti: Wentworth + Moranbah Airport)	giornaliera	2011-2019	pacchetto in R (bomrang), Sparks et al., 2017	
Air mass cluster (120-h backwards trajectories every 6 h - 00, 06, 12, 15 UTC - at a height of 100 m above ground level)	puntuale (Moranbah station)	ogni 6 ore	2011-2019	NCEP/NCAR reanalisi + HYSPLIT trajectory model	

*nd = non disponibile

Sono stati utilizzati diversi parametri meteorologici, sia da stazioni di misura che dal database ERA5. In particolare:

- Velocità e direzione del vento (misurata alla stessa stazione di QA di Moranbah utilizzando un sensore ultrasonico all'altezza di 10 m sul livello del mare);
- Temperatura (misurata alla stessa stazione di QA di Moranbah);
- Dati di piovosità giornaliera elaborati con il pacchetto in R *bomrang* considerando due diversi siti: Wentworth a circa 30 km da Moranbah per il periodo 2011-2019 e l'aeroporto di Moranbah per il periodo 2012-2019. Se lo stesso giorno entrambi i siti fornivano un dato di pioggia, si è considerato il valore medio dei due dati di pioggia, altrimenti se solo una delle stazioni ha misurato un valore di pioggia, è stato assegnato quel valore alla stazione di Moranbah.
- Altezza oraria del boundary layer, contenuto d'acqua nel suolo sui primi 7 cm e pressione sul livello del suolo sono stati considerati dal database European Centre for Medium Weather Forecasting's Reanalysis-5 (ERA5) per il periodo 2011-2019. La media di queste tre variabili sul dominio spaziale ad intervalli orari è stata calcolata e considerata rappresentativa della regione Moranbah.
- Le retrotraiettorie a 120 h calcolate ogni 6 ore ad un'altezza di 100m sul livello del suolo con le reanalisi meteorologiche di NCEP/NCAR e il modello HYSPLIT tra il 2011 e il 2019 usando il pacchetto in R *splitr*.

ii. Dati qualità dell'aria (§ 2.2.1, pag. 3)

- Concentrazioni orarie di PM_{10} misurate alla stazione di Moranbah.
- PM_{10} misurato con TEOM (Tapered Element Oscillating Balance) a 4 m di altezza sul livello del suolo i cui valori sono stati aggiustati secondo il metodo EQPM-1090-079.
- Valori orari di concentrazioni negativi sotto i 0.5 $\mu g/m^3$ sono stati rimossi e tutti gli altri valori negativi di concentrazione sono stati fissati a 0.0 $\mu g/m^3$.

iii. Altri dati

- Serie storica degli incendi dal database Geoscience Australia Sentinel Hotspots.

b. Metodo statistico normalizzazione

i. Descrizione algoritmo (§ 3.3, pagg. 5-6)

- Sono stati utilizzati due differenti algoritmi di machine learning:
    - Gradient boosted regression (GBR): metodo applicato a partire dal pacchetto in R *deweather* che usa il pacchetto *gbm*. Applicate funzioni per determinare interaction depth, learning rate e numero minimo di campioni da considerare per ogni nodo. In particolare è stato utilizzato un train fraction di 0.8 a un bag fraction di 0.5 e il modello è stato fatto girare 10 volte per ogni combinazione di iperparametri.

Iperparametro del modello	Valori
Number of trees	[10, 500, 1000, 1500, 2000, 4000]
Learning rate	[0.1, 0.05, 0.01, 0.005 , 0.001]
Interaction depth	[4, 6, 8, 10]
Minimum samples per node	[10, 14, 18]

La configurazione ottimale è stata determinata con un numero di alberi pari a 1000, un interaction depth pari a 4, a learning rate di 0.005 e un minimo di osservazioni per ogni node pari a 10.

    - Random forest models (RFM): metodo basato sul pacchetto in R *rmweather* che usa il pacchetto in R *ranger*. Il valore di default del training fraction è stato fissato pari a 0.8 e differenti combinazioni di hyper parameters sono state testate. La configurazione ottimale è stata determinata per un numero di alberi pari a 600, 6 variabili per nodo e una dimensione minima del nodo di 10.

Iperparametro del modello	Valori
Number of trees	[100, 200, 400, 600, 800, 1000, 1200, 1400]
Number of variables per node	[2, 4, 6, 8, 10]
Minimum node size	[2, 4, 6, 8]

ii. Validazione del modello (par 3.2, pag. 6)

- Il modello costruito con RFM ha prestazioni migliori del GBR con un coefficiente di determinazione, R^2, che varia da 0.49 a 0.59 nei RFM a 0.25-0.49 nei GBR.
- La configurazione ottimale per RFM e GBR ha mostrato un RMSE pari a 19.5 e 21.1, rispettivamente.

iii. Analisi criticità e punti di forza

- Analizzate dipendenze delle concentrazioni di PM_{10} da ogni singola variabile predittiva.

c. Metodo statistico per la stima del contributo delle misure attuate

i. Descrizione algoritmo

In questo caso il lavoro non è stato applicato per determinare il contributo delle misure di distanziamento sulle concentrazioni. Le tecniche di GBR e RFM sono state applicate per determinare l'influenza di singoli fattori sull'andamento delle concentrazioni di PM_{10} nella località di Moranbah. Per ogni singola variabile si è effettuata un'analisi dei trend e di influenza sulle concentrazioni di PM_{10}, su cui è stata realizzata un'analisi di trend sia sulla serie misurata che dopo normalizzazione. Tutti gli algoritmi sono stati sviluppati in R, oltre ai precedenti anche il pacchetto *openair* per analisi dei trend (metodo Theil-sen). L'articolo rimanda ad un *repository* che però non risulta funzionante, https://github.com/marc-mallet/moranbah_pm10. Lo studio afferma però che le tecniche

messe a punto nella presente analisi potranno essere successivamente utilizzate per stimare l'impatto del COVID-19 sull'andamento della qualità dell'aria in Australia.

ii. <u>Validazione del modello</u>

- Analizzato andamento concentrazione PM_{10} normalizzato in cui sono stati determinati breakpoints (10 date) legati a variazioni emissive più che ad effetti meteorologici/ambientali.

3.5 Scheda analisi bibliografia RF n°5

Vu, T. V, Shi, Z., Cheng, J., Zhang, Q., He, K., Wang, S., & Harrison, R. M. (2019). Assessing the impact of clean air action on air quality trends in Beijing using a machine learning technique. *Atmospheric Chemistry and Physics, 19*, 11303–11314. *https://doi.org/10.5194/acp-19-11303-2019*

Metodo statistico per la normalizzazione e stima del contributo

a. Dati di input

i. Dati meteo

(§2.1, pag 11304; supplement Section S3, pag S5)

Variabili meteo iniziali	selezionate nel modello finale	Risoluzione spaziale	Risoluzione Temporale	Serie storica minima	Potenziali fonti dati	Formato dei dati
wind speed	x	Dati della stazione del Beijing International Airport	h	17/01/2013 - 31/12/2017	worldMet R package	.csv
wind direction	x					
temperature	x					
relative humidity	X					
pressure	x					

ii. Dati qualità dell'aria delle stazioni di monitoraggio

(§ 2.1, pag 11304; supplement Section S1, pag S2)

Inquinanti	Risoluz. Temp.	Serie storica minima	Aggregazione minima dei dati	Classificaz. tipo stazione	Formato dei dati
$PM_{2.5}$	h	17/01/2013 - 31/12/2017		urban, suburban e rural areas (12 national air quality stations) classificate in queste 3 categorie basandosi su un clustering gerarchico	.csv
PM_{10}					
NO_2					
SO_2					
O_3					
CO					

iii. Altri dati

(§ 2.1, pag 11304 e § 2.2.1 pag. 11305)

Variabili esplicative non meteorologiche: week/weekend, mesi da 1 to 12, giorno dell'anno da 1 a 365, ora del giorno da 0 a 23, Unix epoch time (Ttrend).
La variabile ttrend è calcolata come segue:

$$t_{\text{trend}} = \text{year}_i + \frac{t_{\text{JD}} - 1}{N_i} + \frac{t_H}{24N_i},$$

con Ni numero di giorni nell'anno i (dal 2013 al2017), tH ora del giorno(0–23), tJD giorno dell'anno (1–365)) (Carslaw and Taylor, 2009).

I dati mensili di emissione degli inquinanti derivanti dal Multi-resolution Emission Inventory for China sono stati usati non per lo sviluppo della RF ma per confrontare le diminuzioni osservate nelle concentrazioni degli inquinanti, a valle della normalizzazione meteorologica, con le diminuzione delle emissioni.

iv. Analisi criticità e punti di forza

Manca il PBL come variabile esplicativa. Nonostante questo ottengono un buon risultato in termini di fitting del modello. La disponibilità del dato orario del particolato (sia $PM_{2.5}$ che PM_{10}) è un vantaggio rispetto alla disponibilità di punti di campionamento con misura giornaliera.

b. Metodo statistico normalizzazione

i. Descrizione algoritmo

Formulazione: Random Forest per la normalizzazione meteorologica delle serie dei dati e, sui dati normalizzati a livello meteorologico, test di Theil-Sen per la valutazione del trend. Il dataset di training comprendeva il 70% delle osservazioni, con le restanti invece utilizzate per la fase di test.
Rispetto all'algoritmo utilizzato da Grange, per lo sviluppo della RF, hanno utilizzato il ricampionamento delle sole variabili meteo (§2.2.2, pag. 11306).

I pacchetti R utilizzati sono:

- Random-ForestExplainer
- normalweatherr e rmweather (per la costruzione della RF)
- openair (per lo stimatore Theil-Sen)

Nell'articolo sono descritte le procedure utilizzate ed è reso disponibile il codice sorgente.

ii. Validazione del modello

Gli indicatori di performance della RF utilizzati sono di seguito riportati: RMSE root-mean-square error FAC2 (fraction of predictions with a factor of two), MB (mean bias), MGE (mean gross error), NMB (normalised mean bias), NMGE (normalised mean gross error), COE (Coefficient of Efficiency), IOA (Index of Agreement).

iii. Analisi criticità e punti di forza

Non è spiegato come fanno l'aggregazione spaziale tra i risultati ottenuti nelle diverse stazioni.
L'adattamento statistico del modello risulta buono per tutti gli inquinanti. La parte metodologica dell'articolo è ben dettagliata, con materiale di approfondimento e codici condivisi.

c. Metodo statistico per la stima del contributo delle misure attuate

i. Descrizione algoritmo

(§3.2 pag 11307-11308)
In questo articolo si è voluto valutare l'efficacia delle misure attuate nella regione di Beijing nel quinquennio 2013-2017 per la riduzione dell'inquinamento atmosferico. Si è quindi valutato un decremento delle concentrazioni degli

inquinanti. A tale scopo è stata applicata una RF sulle serie dati dei diversi inquinanti per ottenere una normalizzazione meteorologica dei valori delle concentrazioni. A valle di questo processo, si sono quindi confrontate le concentrazioni degli inquinanti nel 2013 (all'inizio del piano quinquennale delle misure) e nel 2017 e si è valutato il decremento. Si è calcolato anche il numero di giorni in cui i livelli di $PM_{2.5}$ sono risultati superiori ai 75µg/m^3.

ii. Validazione del modello (par. 3.3, pagg. 11308 – 11309)

Il decremento delle concentrazioni degli inquinanti ottenuto sui dati normalizzati meteorologicamente è stato confrontato (su medie mensili) con il decremento stimato dal modello di dispersione delle emissioni WRF-CMAQ per valutare la convergenza dei diversi approcci. Il valore previsto dai due modelli per il $PM_{2.5}$ del 2017 è stato molto simile (61 µg/m^3 per la RF e 61.8 e 62.4 µg/m^3 per il WRF-CMQA). Sono state confrontate anche le incertezze dei due modelli considerando le MEDIE MENSILI previste e osservate del $PM_{2.5}$ in termini di coefficiente di correlazione (r= 0.82 per il WRF-CMAQ e r=0.99 per la RF) e in termini di differenza percentuale (compresa tra il 3% e il 33.6% per WRF-CMAQ e 0.4% e 7.9% per la RF).

3.6 Sintesi modelli Random Forest

Nelle applicazioni dei modelli RF descritte negli articoli di *Grange et al*. per le concentrazioni degli inquinanti considerati (PM_{10}, NO_2, NO_x e SO_2) è stata stimata la tendenza normalizzata meteorologicamente. Adoperando più di mille previsioni è stata ottenuta però solo una minima riduzione del rumore. La formazione dei modelli è stata condotta sull'80% dei dati di input e l'altro 20% è stato invece utilizzato per validare i modelli.

Le stime delle tendenze sono state valutate con il test Theil-Sen, mentre attraverso l'utilizzo del metodo di identificazione dei breakpoint (strucchange R package) si è tentato di investigare gli effetti sulla qualità dell'aria degli interventi messi in atto per ridurre le emissioni di tali inquinanti.

La RF non offre la possibilità di determinare l'errore o l'incertezza delle stime direttamente, allora vengono fatti crescere dei modelli RF con set di input campionati in modo casuale (bootstrap) per poi valutare l'importanza delle variabili indipendenti.

In altre applicazioni (Petetin et al. 2020) è stato utilizzato il metodo del Gradiente Boosting Machine (GBM), decision tree-based ensemble appartenente alla boosting family, per stimare la concentrazione di NO_2 "business as usual" (espressa in ppbv) che sarebbe stata osservata in assenza del lockdown (avvenuto anche in Spagna nella primavera dell'anno 2020), per stazioni di tipologia urbana e da traffico. Il modello è stato addestrato sui tre anni completi (2017-2019) e poi utilizzato per predire i le concentrazioni business-as-usual di NO_2 per i seguenti 4 mesi del 2020. È stato riscontrato che i modelli predittivi di ML hanno funzionato molto bene nella maggior parte delle località. Durante il periodo di studio, è stato stimato che le misure attuate durante il periodo di lockdown abbiano comportato una riduzione media del 50% dei livelli di NO_2 su tutte le province e le isole. Ampliando il periodo di addestramento a 4 o 5 anni non sono stati notati miglioramenti. Per la validazione dei risultati sono state utilizzate le seguenti metriche: mean bias (MB), normalized mean bias (nMB), root mean square errore (RMSE), normalized root mean square error (nRMSE) e Person correlation coefficient (PCC).

Nel lavoro di Kamiska et al. 2018, la RF vengono utilizzate per modellare le relazioni di regressione tra le concentrazioni degli inquinanti NO_2, NOx e $PM_{2,5}$ e nove variabili descrittive delle condizioni meteorologiche, delle condizioni temporali e flusso del traffico, nel biennio 2015 e 2016. Il modello costruito è stato addestrato su tutti e nove i periodi considerati utilizzando come training set il 50% di tutto il campione e il 30% del campione per il test. Il processo di apprendimento (learning process) si è fermato quando per 10 cicli l'errore era inferiore al 5%.
L'importanza delle variabili predittive è stata determinata come somma, per ogni nodo dell'albero, degli incrementi del paramento ΔR (resubstitution estimate).
È stato riscontrato che sia l'adattamento che l'importanza di particolari predittori dipendono dalla stagione. Il migliore fit è stato ottenuto per i modelli sviluppati per la stagione calda (periodo di sei mesi da aprile a settembre) e per la stagione estiva (giugno e agosto). La variabile esplicativa più importante nei modelli per gli ossidi di azoto è stata il flusso del traffico, mentre nel caso del $PM_{2,5}$ le più importanti sono state le condizioni meteorologiche, in particolare temperatura, velocità e direzione del vento. Variabili

temporali (ad eccezione del *mese* nel caso del $PM_{2,5}$) non hanno avuto effetti significativi sulle concentrazioni di gli inquinanti studiati.

Nel lavoro di Mallet 2020, le tecniche di GBR e RFM sono state applicate per determinare l'influenza di singoli fattori (emissioni locali, presenza di miniere di carbone a cielo aperto, diminuzione del contenuto di acqua del suolo nella regione circostante, che può facilitare maggiori emissioni di polveri) sull'andamento delle concentrazioni di PM_{10} nella località di Moranbah (Queensland, Australia). Per ogni singola variabile è stata effettuata un'analisi dei trend delle concentrazioni di PM_{10}, sia sulla serie misurata che su quella normalizzata. Lo studio afferma che le tecniche messe a punto potranno essere successivamente utilizzate per stimare l'impatto del COVID-19 sull'andamento della qualità dell'aria in Australia.

Il modello costruito con RFM ha prestazioni migliori del GBR con un coefficiente di determinazione, R^2, che varia da 0.49 a 0.59 nei RFM a 0.25-0.49 nei GBR. La configurazione ottimale per RFM e GBR ha mostrato un RMSE pari a 19.5 e 21.1, rispettivamente.

Nel lavoro di Vu et al., 2019, si è voluto valutare l'efficacia delle misure attuate nella regione di Beijing nel quinquennio 2013-2017 per la riduzione dell'inquinamento. A tale scopo è stato applicato un modello RF sui dati di diversi inquinanti per ottenere una normalizzazione meteorologica dei valori delle concentrazioni. Si sono quindi confrontate le concentrazioni così ottenute degli inquinanti nel 2013 (all'inizio del piano quinquennale delle misure) e nel 2017 e si è valutato il decremento.

Rispetto all'algoritmo utilizzato da Grange, per lo sviluppo della RF, è stato utilizzato il ricampionamento delle sole variabili meteo (§2.2.2, pag. 11306). Il decremento delle concentrazioni degli inquinanti ottenuto sui dati normalizzati a livello meteorologico è stato confrontato (su medie mensili) con il decremento stimato dal modello di dispersione degli inquinanti, WRF-CMAQ, per valutare la convergenza dei diversi approcci. Il valore previsto dai due modelli per il $PM_{2.5}$ del 2017 è stato molto simile (61 $\mu g/m^3$ per la RF e 61.8 e 62.4 $\mu g/m^3$ per il WRF-CMQA).

4 TEST MULTIPLI NELL'ANALISI SPAZIO TEMPORALE DEI DATI AMBIENTALI

Cortés, J., Mahecha, M., Reichstein, M., Brenning, A. (2020). Accounting for multiple testing in the analysis of spatio-temporal environmental data. *Environmental and Ecological Statistics*, 27, 293–318.

https://doi.org/10.1007/s10651-020-00446-4.

L'articolo affronta il problema dei test multipli, che si è posto in maniera urgente con l'elaborazione dei dati provenienti dal remote sensing e dall'Earth system simulation, in cui vengono analizzati i dati di una griglia, con migliaia o anche milioni di celle, per ognuna delle quali deve essere vagliata una ipotesi. Qualora l'ipotesi nulla che viene testata sia vera a livello globale (sull'intera griglia), per come è definito il concetto di significatività, ci si aspetta che un numero di celle pari ad α diano luogo a falsi positivi (ipotesi nulla rifiutata anche se in realtà vera); inoltre, dati spazialmente autocorrelati possono dare origine a rifiuti dell'ipotesi nulla clusterizzati, il che può essere fuorviante in un'analisi dei modelli spaziali. La probabilità di ottenere almeno un falso positivo in una "famiglia" di test è chiamata Familiwise Error Rate (FWER) ed è una funzione crescente del numero dei test che vengono effettuati tale che con già un centinaio di test c'è una probabilità >del 99% di ottenere almeno un falso positivo.
Due strategie sono comunemente utilizzate per controllare il FWER. Una consiste nel definire una nuova soglia di significatività che tenga conto del numero dei test che vengono effettuati (i metodi che si basano su questo principio vengo definiti "Bonferoni related") mentre l'altra consiste nel definire una soglia di significatività usando la distribuzione del "maximum statistic". L'articolo introduce metodi basati sulla distribuzione del "maximum statistic" e li confronta con i metodi "Bonferoni related" nel contesto dei dati ambientali geospaziali. La performance dei diversi metodi viene valutata con uno studio su dati simulati e su due dataset di dati reali.
Tra i metodi "Bonferoni related" vengono ricordati:
- la **correzione di Bonferroni**: stabilisce una nuova soglia di significatività dividendo α per il numero M di celle;
- il **metodo di Walker** che usa la distribuzione del p-value minimo – che come è noto segue una distribuzione beta – per stabilire la soglia di significatività;
- i metodi con una soglia di significatività variabile: in questo caso si considerano i test uno per volta, a partire da quello che ha dato origine al p-value più alto (**metodo step-up Hochenberg**) o più basso

(**metodo stepdown, Holm**) e la soglia di significatività α è divisa per 1, 2, 3, … (metodo step-up) o per M, M-1, M-2, … (metodo step-down); ci si ferma quando un test risulta significativo e si dichiara significativo quel test e tutti quelli con p-value inferiore;

- i metodi che controllano il **False Detection Rate (FDR)**, ovvero la proporzione di ipotesi nulle respinte erroneamente rispetto al numero totale di ipotesi nulle respinte. Un piccolo FDR garantisce che le "scoperte" (casi in cui si rifiuta l'ipotesi nulla) siano affidabili, senza imporre vincoli sulla probabilità di fare almeno una falsa "scoperta". Questo metodo è stato proposto da **Benjamini and Hochberg, BH**). **Benjamini and Yakuteli (BY)** hanno modificato il metodo BH che controlla il FDR sotto l'ipotesi di autocorrelazione positiva (o di indipendenza), per tener conto di altri casi di dipendenza tra le statistiche-test.

Tutti questi metodi Bonferroni-related, però, per M grandi differiscono poco da Bonferroni, presentando lo stesso problema di essere troppo conservativi, comportando difficoltà ad identificare celle significative.

I metodi "maximum distribution":

Comprendono 2 tipologie di test: uno per la maximum distribution della statistica test (in seguito chiamato **maxT**) e uno per la maximum distribution della dimensione del cluster sopra la soglia della statistica test (**STCS**). Si tratta di test non parametrici di permutazione: viene calcolata la statistica test sotto ogni possibile permutazione dei dati (se le permutazioni possibili sono troppe, ci si può anche limitare ad un sottoinsieme). Per questo calcolo si assume (ipotesi nulla) che i dati siano interscambiabili, ovvero che non conti l'ordine. Ad es. se si ha a che fare con una serie storica e si effettua un test di mann-Kendall per l'analisi di un trend assumendo che i dati siano intercambiabili, in questo caso si sta ipotizzando che non ci sia alcun trend. Il p-value del test effettuato sulla serie storica originale è dato dalla proporzione del numero delle permutazioni della serie storica che danno luogo a statistiche-test più grandi o uguali a quella osservata sulla serie storica originale. Permutare l'intera griglia conserva l'autocorrelazione spaziale presente nei dati.

Si è fatto l'esempio di un test di Mann-Kendal per testare la presenza di un trend, ma può essere utilizzata una qualunque statistica test valida.

Questo metodo ha il vantaggio di essere meno conservativo rispetto ai Bonferroni-related test. Si pensi che con una griglia di solo 10000 celle, la soglia di Bonferroni con un livello di significatività α pari a 0.05 diventa = 0.05/10000= $5*10^{-6}$; con 1000 permutazioni, il minimo valore di p che può essere raggiunto è 1/1000, cioè $1*10^{-3}$, che è parecchi ordini di grandezza maggiore.

La maximum distribution della dimensione del cluster sopra la soglia (**STCS**) si basa sul numero di celle della griglia significative che sono adiacenti. Per entrambi i metodi, si permuta l'intera immagine contemporaneamente e si ricalcola la statistica test in ogni cella della griglia. Col metodo maxT, si registra la statistica test massima tra tutte le celle della griglia; col metodo STCS, si registra la dimensione del più grande cluster di rete significativo. Quindi si ripetono questi passaggi N volte. La distribuzione della maximum statistic è formata dalla statistiche registrate ad ogni permutazione. Entrambi questi metodi permettono di controllare il FWER alla significatività α_{global} desiderata, settando la soglia di significatività u al 100*(1- α_{global})esimo percentile delle relative distribuzioni massime. Per maxT, è dichiarato significativo il test in ogni cella della griglia in cui il cui valore assoluto della statistica test supera u; per STCS, u è definito in termini di numerosità del cluster, per cui ogni cluster più grande di u è dichiarato significativo.

I metodi vengono testati su un set di dati simulati (su cui sono stati imposti trend e funzioni di autocorrelazioni noti) e su due dataset reali.

I metodi Bonferroni-related raggiungono un FWER ben al di sotto del livello nominale, il che influisce sulla loro potenza di test globale (ovvero che valuta un'ipotesi sull'intera griglia e non su ogni singola cella) e sulla loro potenza all'interno della griglia.

Solo maxT controlla il FWER al livello di significatività globale α_{global} desiderato. Il STCS è leggermente più conservativo (0.02-0.03) per bassi livelli di autocorrelazione ma si avvicina al livello nominale all'aumentare dell'autocorrelazione spaziale. I metodi Bonferroni-related sono sovrapponibili e raggiungono un FWER di 0.01, indipendentemente dal grado di autocorrelazione spaziale. BY risulta il più conservativo, con un FWER di 0.0002.

I metodi con permutazione presentati costituiscono una valida alternativa per l'indirizzamento del problema dei test multipli nelle scienze ambientali. Specificatamente, il metodo STCS controlla il FWER e supera tutti gli altri metodi sia come potenza del test globale, sia nella potenza all'interno della griglia in tutti gli scenari tranne uno, quando c'è una piccola area con un effetto (0,25%) e una forte correlazione spaziale ($\geq$ 0,8). Che il metodo STCS fallisca in questo caso non è inaspettato. Il clustering infatti prende in considerazione solo la dimensione del cluster, non i singoli p-value, perciò, con una autocorrelazione spaziale più forte, i cluster che appaiono casualmente diventano più grandi della dimensione del cluster con il trend. Questa situazione rende i cluster con trend invisibili al metodo STCS.

In termini di potenza del test all'interno dell'immagine, il metodo di BH identifica più segnali rispetto al metodo maxT, ma a questo vantaggio corrisponde un costo in termini di un maggior numero di falsi

positivi e nessun potere di localizzazione: le singole celle della griglia non possono essere dichiarate significative. Il metodo maxT identifica meno celle della griglia, ma permette di individuare singole celle della griglia significative. Questa caratteristica può essere di importanza critica nell'analisi dei dati geospaziali. Non correggere per test multipli porterebbe ad individuare quasi sempre una significatività di campo e alla possibile errata interpretazione di pattern spaziali spuri.

Dal momento che la distribuzione massima di una statistica di test può essere derivata per qualsiasi metodo di permutazione, è semplice tenere sotto controllo il FWER con i metodi presentati in questo paper. Tuttavia, esistono limitazioni per l'uso dei metodi di permutazione. Ci sono casi in cui una procedura di permutazione non è semplice o è impossibile: per esempio, se la statistica del test è invariante alle permutazioni (come ad esempio, il test t su un campione). In casi simili il metodo BH costituisce una valida alternativa.

Sebbene siano stati sviluppati molti metodi per test multipli, da quando Livezey e Chen (1983) hanno introdotto il loro metodo nelle scienze ambientali, questi non tengono conto dell'autocorrelazione spaziale, che invece è spesso presente nei dati ambientali. I metodi di permutazione introdotti in questo paper catturano l'autocorrelazione spaziale nella distribuzione maximum-statistic e in questo modo migliorano i metodi precedenti: i metodi di permutazione hanno una potenza globale maggiore rispetto a tutti gli altri metodi qui confrontati, incluso il metodo BH.

4.1 Esempio di applicazione della correzione di Bonferroni

Connerton, P., de Assunção, J.V., de Miranda, R.M., Slovic, A.D., Pérez-Martínez, P.J., Ribeiro, H. (2020). Air Quality during COVID-19 in Four Megacities: Lessons and Challenges for Public Health. *International Journal of Environmental Research and Public Health, 17(14):5067. https://doi.org/10.3390/ijerph17145067.*

Lo studio descritto in questo articolo analizza gli effetti della quarantena e delle politiche di distanziamento sociale, attuate a causa della pandemia di Coronavirus Disease 2019 (COVID-19), sulla qualità dell'aria e le relative possibili conseguenze sulla salute umana, in quattro megalopoli occidentali: San Paolo in Brasile; Parigi in Francia; e Los Angeles e New York negli Stati Uniti.

Nella tabella seguente sono riportati, il periodo in studio, le variabili meteorologiche e gli inquinanti atmosferici esaminati, e infine le statistiche descrittive elaborate sia per le variabili meteorologiche che per gli inquinanti, nei due periodi, marzo 2015-2019 e marzo 2020.

Periodo esaminato	Variabili meteorologiche	Inquinanti atmosferici	Statistiche descrittive	
marzo 2015-2020	temperatura (T) umidità relativa (RH) velocità del vento (WS) precipitazioni cumulative (P)	monossido di carbonio (CO) ozono (O_3) particolato fine ($PM_{2.5}$) biossido di azoto (NO_2)	media annuale deviazione standard numero dati mediana	marzo 2015-2019 marzo 2020

Le statistiche descrittive sono state confrontate e fittate con un modello lineare generale (GLM) utilizzando CO, O_3, NO_2 e $PM_{2.5}$ come variabili dipendenti e i parametri climatici a lungo termine - temperatura, umidità relativa e velocità del vento - come variabili esplicative/indipendenti.

Il GLM ha stimato la qualità dell'aria durante il giorno t, nel mese di marzo negli anni 2015-2020, utilizzando i dati di qualità dell'aria rilevati dalle reti regionali.

Il modello di regressione utilizzato è il seguente:

$$Air\_quality_{i,t\ (pollutant\ i,\ march\ day\ t)} = a_0 + a_1 P_t + W'_t\ a_2^W$$

dove Pt è una variabile *dummy* (uguale a 1 per l'anno 2020 e uguale a 0 per gli anni 2015–2019), che rappresenta l'effetto relativo delle limitazioni delle attività umane nel mese marzo 2020 rispetto agli anni precedenti; a_0 (intercetta) e a_1 (effetto scenario COVID-19) sono coefficienti di regressione ottenuti dai minimi quadrati ordinari (OLS); W'_t è il vettore dei dati climatici che possono avere un impatto sull'inquinamento atmosferico, e a_2^W sono i tre coefficienti di regressione relativi a temperatura dell'aria (°C), umidità relativa (%) e velocità del vento (m/s).

Infine, utilizzando l'analisi della varianza e il test post-hoc di Bonferroni è stato effettuato un confronto fra le statistiche descrittive relative ai due scenari, quello precedente (2015-2019) e quello interessato dal COVID-19 (2020).

Di seguito si riportano le riduzioni (%) delle concentrazioni di CO, NO_2 e $PM_{2.5}$ nel marzo 2020, rispetto al quinquennio precedente 2015-2019:

	Riduzione (%) delle concentrazioni		
	CO	**NO_2**	**$PM_{2,5}$**
San Paolo	40%	25%	12%
Los Angeles	24%	38%	37%
New York	19%	24%	24%
Parigi	67%	39%	28%

Riguardo all'ozono, è stata rilevata una diminuzione delle concentrazioni solo a Los Angeles; nelle restanti città è risultato un aumento delle concentrazioni nel periodo considerato.

Dallo studio è risultato che l'effetto delle restrizioni imposte e, in particolare della riduzione del traffico, sul miglioramento della qualità dell'aria è stato più rilevante di quello della meteorologia.

Dai risultati di questo studio, è emerso che le misure di distanziamento sociale, oltre ad abbassare il rischio di trasmissione del virus, hanno avuto anche un effetto positivo sulla qualità dell'aria, riducendo potenzialmente le malattie respiratorie e cardiovascolari associate all'esposizione agli inquinanti atmosferici nel periodo in esame.

BIBLIOGRAFIA

Barmpadimos, I., Hueglin, C., Keller, J., Henne, S., & Prévôt, A. S. H. (2011). *Influence of meteorology on PM 10 trends and variability in Switzerland from 1991 to 2008*. Atmospheric Chemistry and Physics, *11*, 1813–1835. https://doi.org/10.5194/acp-11-1813-2011.

Cameletti, M. (2020). The Effect of Corona Virus Lockdown on Air Pollution: Evidence from the City of Brescia in Lombardia Region (Italy). Atmospheric Environment, 239:117794. https://doi.org/10.1016/j.atmosenv.2020.117794.

Carslaw, D. C., & Carslaw, N. (2007). *Detecting and characterising small changes in urban nitrogen dioxide concentrations*. *Atmospheric Environment, 41*(22), 4723–4733. https://doi.org/10.1016/j.atmosenv.2007.03.034

Carslaw, D. C., Beevers, S. D., & Tate, J. E. (2007). *Modelling and assessing trends in traffic-related emissions using a generalised additive modelling approach*. *Atmospheric Environment, 41*, 5289–5299. https://doi.org/10.1016/j.atmosenv.2007.02.032.

Cortés, J., Mahecha, M., Reichstein, M., Brenning, A. (2020). *Accounting for multiple testing in the analysis of spatio-temporal environmental data*. Environmental and Ecological Statistics, *27*, 293–318. https://doi.org/10.1007/s10651-020-00446-4.

Connerton, P., de Assunção, J.V., de Miranda, R.M., Slovic, A.D., Pérez-Martínez, P.J., Ribeiro, H. (2020). *Air Quality during COVID-19 in Four Megacities: Lessons and Challenges for Public Health*. International Journal of Environmental Research and Public Health, *17*, 5067.

Grange, S. K., & Carslaw, D. C. (2018). *Using meteorological normalisation to detect interventions in air quality time series Science of the Total Environment Using meteorological normalisation to detect interventions in air quality time series*. Science of the Total Environment, 653(November), 578–588. https://doi.org/10.1016/j.scitotenv.2018.10.344.

Grange, S. K., Carslaw, D. C., Lewis, A. C., Boleti, E., & Hueglin, C. (2018). *Random forest meteorological normalisation models for Swiss PM 10 trend analysis*. Atmospheric Chemistry and Physics, 18, 6223–6239. https://doi.org/10.5194/acp-18-6223-2018

Kaminska, J.A. (2018). *The use of random forests in modelling short-term air pollution effects based on traffic and meteorological conditions: A case study in Wroclaw.* Journal of Environmental Management, *217*, 164-174. ISSN 0301-4797. https://doi.org/10.1016/j.jenvman.2018.03.094.

Mallet, M.D. *Meteorological normalisation of PM_{10} using machine learning reveals distinct increases of nearby source emissions in the Australian mining town of Moranbah*. Atmospheric Pollution Research, in press (2020).

Ordóñez, C., Garrido-Perez, J. M., & García-Herrera, R. (2020). *Early spring near-surface ozone in Europe during the COVID-19 shutdown: Meteorological effects outweigh emission changes*. Science of the Total Environment, 747(December 2019). https://doi.org/10.1016/j.scitotenv.2020.141322

Petetin, H., Bowdalo, D., Soret, A., Guevara, M., Jorba, O., Serradell, K., & Pérez García-Pando, C.P. (2020). *Meteorology-normalized impact of COVID-19 lockdown upon NO_2 pollution in Spain.* Atmospheric Chemistry and Physics Discussions, 20, 11119–11141,. https://doi.org/10.5194/acp-20-11119-2020.

Vu, T. V, Shi, Z., Cheng, J., Zhang, Q., He, K., Wang, S., & Harrison, R. M. (2019). *Assessing the impact of clean air action on air quality trends in Beijing using a machine learning technique*. Atmospheric Chemistry and Physics, *19*, 11303–11314. *https://doi.org/10.5194/acp-19-11303-2019*

Wood S.N. (2017), *Generalized Additive Models An Introduction with R second edition*, Chapman & Hall Book, ISBN 13: 978-1-4987-2833-1.

Xiang, J., Austin, E., Gould, T., Larson, T., Shirai, J., Liu, Y., Marshall, J., & Seto, E. (2020). *Impacts of the COVID-19 responses on traffic-related air pollution in a Northwestern US city*. Science of the Total Environment, 747:141325. https://doi.org/10.1016/j.scitotenv.2020.141325

Zuur A.F. et alii (2009), *Mixed Effects Models and Extensions in Ecology with R*, Springer, ISBN 978-0-387-87457-9.

Zuur A.F. (2012), *Beginner's Guide to Generalized Additive Models with R*, Highland Statistics Ltd.