Università degli studi di Bergamo

Corso di laurea magistrale in Ingegneria Gestionale e Informatica

Scuola di Ingegneria

Statistica II - Modelli dinamici e previsione statistica

Analisi delle emissioni di ammoniaca in territorio lombardo (2000-2020)

Autori

Lorenzo Conti

Samuele Ferri

Fabio Sangregorio

Matteo Tufilli

11 Gennaio 2021

Indice

Capitolo 1

Introduzione

In questo elaborato presentiamo un'analisi dell'andamento temporale delle emissioni di NH_3 legate all'attività agricola su territorio lombardo, esaminando i dati forniti dal *Copernicus Atmosphere Monitoring Service* (CAMS)[1].
Il dataset si presenta in formato NetCDF e si compone di dati con frequenza mensile in un arco temporale che va dal 2000 al 2020, con risoluzione spaziale 0.1×0.1 gradi.
L'unità di misura per le emissioni di NH_3 è il *tera-grams* (10^{12} grammi) per anno (Tg/yr).

Figura 1.1: Logo del Copernicus Atmosphere Monitoring Service (CAMS)

I dati sono basati sulle emissioni annuali 2000-2012 *EDGARv4.3.2*[2] su cui sono stati applicati dei profili temporali mensili provenienti da *CAMS-GLOB-TEMPO*[3]. Dopo il 2012, le emissioni mensili sono state estrapolate fino al 2020 usando un fit lineare degli anni 2011-2014 proveniente da *CEDS Global Inventory*[4]. Quindi i dati sono emulati, non stocastici.

[1] CAMS Global Anthropogenic Emissions (https://eccad3.sedoo.fr/#CAMS-GLOB-ANT)
[2] EDGARv4.3.2 (https://edgar.jrc.ec.europa.eu/overview.php?v=432)
[3] CAMS-GLOB-TEMPO (https://essd.copernicus.org/preprints/essd-2020-175/)
[4] CEDS Global Inventory (http://www.globalchange.umd.edu/CEDS/)

In primo luogo, i dati sono soggetti ad una fase di pre-processing in cui vengono selezionati esclusivamente quelli interni alla regione Lombardia, nonché quelli di interesse. La griglia delle osservazioni risulta quindi quella presentata in Figura 1.2

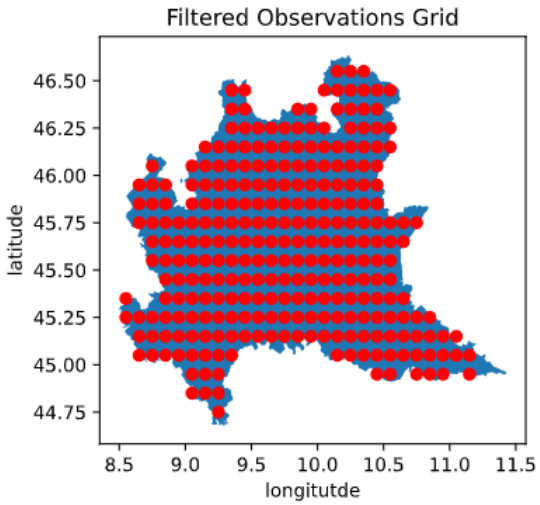

Figura 1.2: Griglia delle osservazioni

In particolare, analizzando il valore medio delle serie storiche nei punti in cui sono presenti le osservazioni, è possibile notare come le emissioni di NH_3 siano maggiori nelle zone di pianura in cui è presente una maggiore attività agricola.

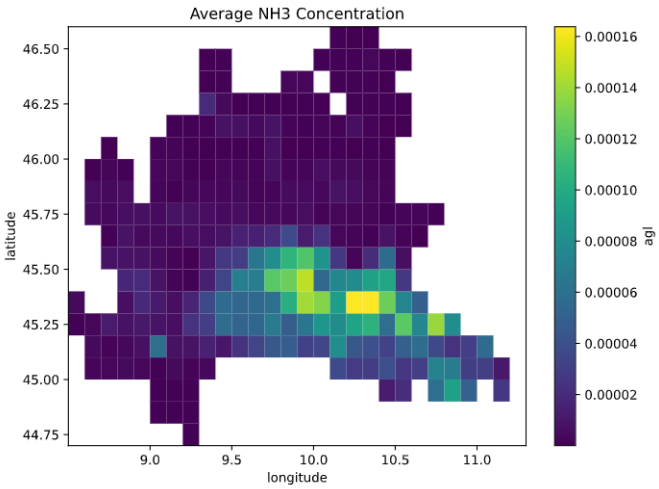

Figura 1.3: Valor medio delle emissioni di NH_3 in Lombardia dal 2000 al 2020

Per avere un'idea dell'area rurale presente in Lombardia si può guardare la figura Figura 1.4 e concentrarsi sulle zone blu.

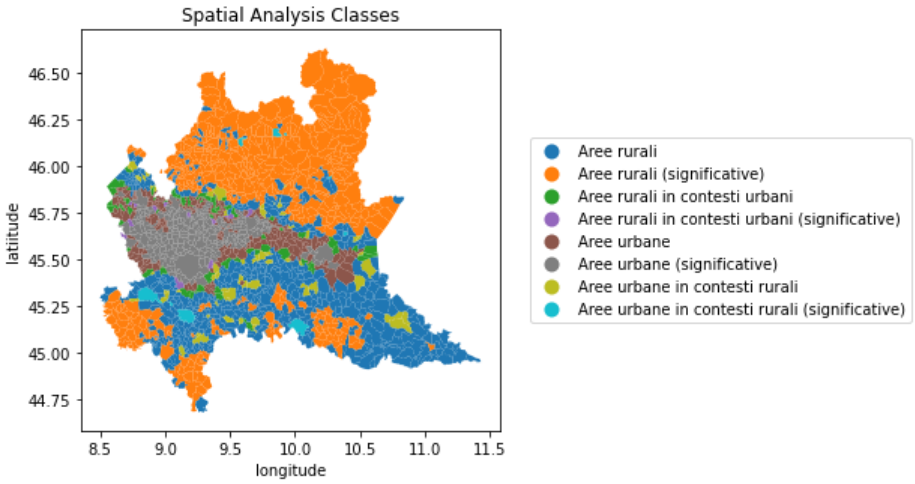

Figura 1.4: Classificazione dei comuni

Sarà quindi di particolare interesse capire se l'andamento delle emissioni è differente nelle varie zone della regione Lombardia e verificarne la similarità o le differenze, ed eventualmente le possibili motivazioni.
Nel corso dell'eleborato si presenterà una analisi più approfondita delle serie storiche individuate sul territorio, seguita dalla presentazione delle tecniche di selezione del modello adottate e dai criteri di selezione del miglior modello in funzione dei risultati ottenuti.
Per ogni tematica, si evidenzieranno le motivazioni che hanno comportato ciascuna decisione ed i problemi riscontrati, con eventuale soluzione.
Il fine ultimo, è quello di individuare un modello matematico per serie storiche in grado di poter prevedere con successo e con elevata precisione l'andamento futuro delle emissioni di ammoniaca.

Capitolo 2

Serie storica

2.1 Analisi preliminare dei dati

Una volta eseguito il preprocessing, il primo step è stato eseguire un'analisi qualitativa della serie storica, in modo da intuirne l'andamento sia in termini di serie storica che in termini spaziali nel territorio della regione (codice visibile nel file `/notebooks/preliminar-analysis/nh3-qualitative-analysis.ipynb`)

La prima visualizzazione effettuata è stata l'andamento di tutte le serie storiche contenute nel file NetCDF, per la durata di un anno (da settembre 2019 ad agosto 2020), in modo da comprendere l'andamento spaziale delle diverse serie storiche. Come mostrato in Figura 2.1, anche se i vari comportamenti sono simili in termini di andamento stagionale, esse risultano eterogenee in termini di curva e scala.

In Figura 2.2 è invece mostrato l'andamento delle serie storiche di un campione di punti spaziali casuali, il quale evidenzia la differenza di scala nelle emissioni.

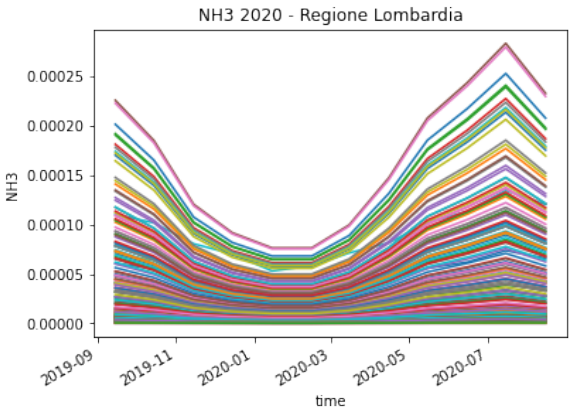

Figura 2.1: Andamento delle serie storiche nell'anno 2020 in ogni punto della regione

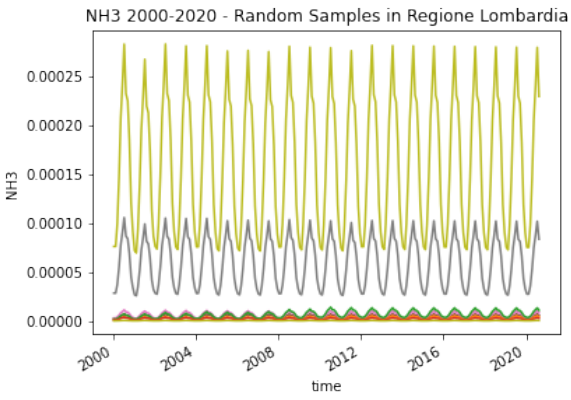

Figura 2.2: Andamento della serie storica di un campione di punti casuali

Per visualizzare meglio questa eterogeneità è stato successivamente mostrato l'andamento annuale medio dell'inquinante nel territorio (mostrato in Figura 2.3). Da qui si evincono due dimensioni dei dati:

- Dimensione spaziale: si nota che la zona territoriale corrispondente alla Pianura Padana rivela emissioni maggiori di ammoniaca. Questo porta a chiedersi se i territori rurali maggiormente presenti in quelle zone possano essere un possibile regressore.
- Dimensione temporale: si vede che le emissioni sono significativamente incrementate nei mesi estivi. Questo indica che verosimilmente la temperatura influisce sull'intensità di emissioni.

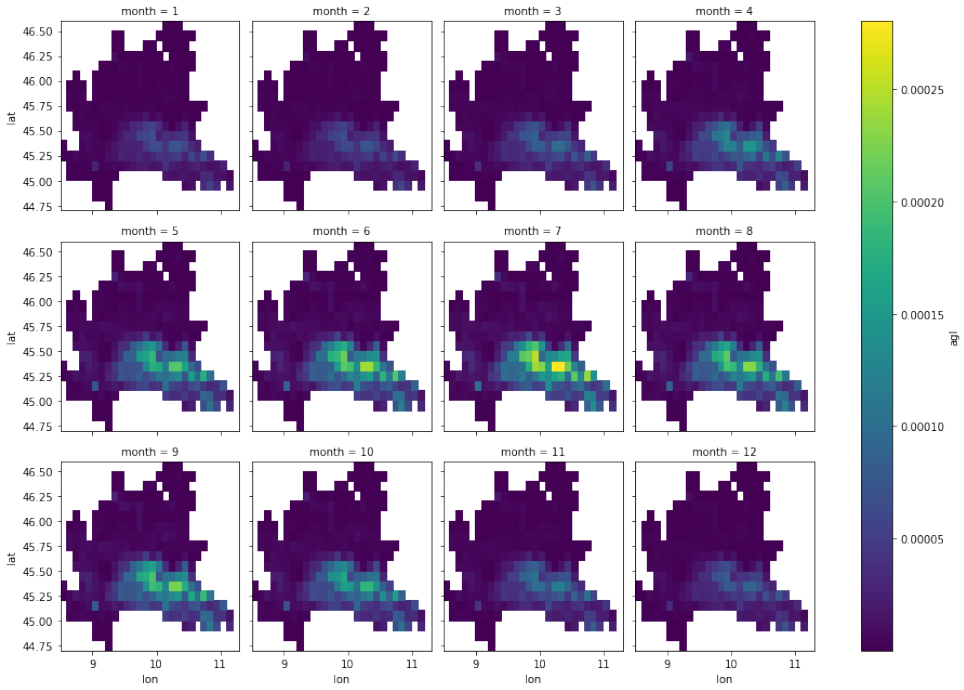

Figura 2.3: Andamento annuale medio dell'ammoniaca nella regione

Per gestire le differenze di scala, si sono dapprima individuati il punto di minimo, medio e massimo nei dati del territorio (visibili come punti di colore rosso in Figura 2.4). Essi sono stati poi successivamente classificati in tre classi sulla base dell'intensità delle emissioni: basse, medie e alte (rispettivamente viola, verde e giallo in Figura 2.4). Nei tre punti rappresentati delle classi si mostra nuovamente che gli andamenti (evidenziati in Figura 2.5) sono simili nonostante le scale molto differenti. Di conseguenza, una domanda interessante è quella di capire se sia possibile individuare un unico modello per eseguire previsioni in ognuno di questi punti, o se al contrario siano necessari modelli diversi.

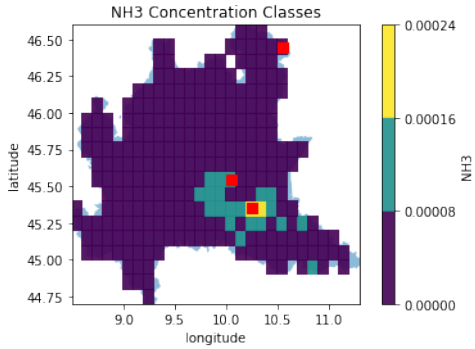

Figura 2.4: Classificazione in tre gruppi sulla base dell'intensità delle emissioni

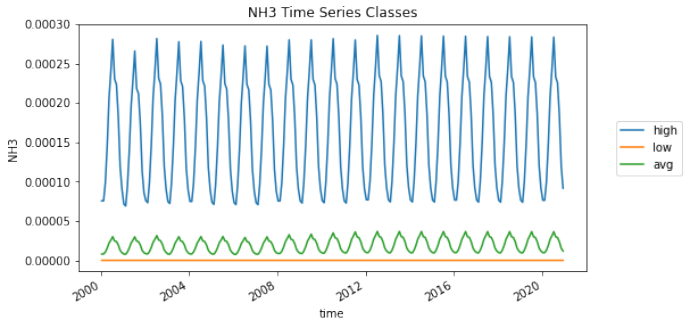

Figura 2.5: Andamento della serie storica nei punti di massimo, minimo e medio

In fase finale dell'analisi preliminare ci si è chiesti se il lockdown potesse avere avuto una qualche influenza sull'intensità delle emissioni, in modo da indagare se ci fosse stata la necessità di eseguire *intervention analysis*. Per la valutazione si è divisa la serie storica negli anni prima del lockdown e nei mesi successivi ad esso, e si sono quindi graficati gli andamenti annuali. Però, come evidenziato in Figura 2.6, l'andamento dell'ultimo anno non presenta anomalie rispetto agli anni precedenti, anzi segue quasi perfettamente i suoi predecessori. Questo ha giustificato il non investigare ulteriormente con un'analisi più approfondita. Una domanda ancora aperta è se la causa di questa uniformità sia il fatto che il lockdown abbia effettivamente avuto impatto nullo sull'intensità delle emissioni, o se sia causata dal fatto che i dati a disposizione non sono dati di misura da sensori terreni, ma bensì dati ricostruiti da modelli, i quali possibilmente non prendono in considerazione il lockdown.

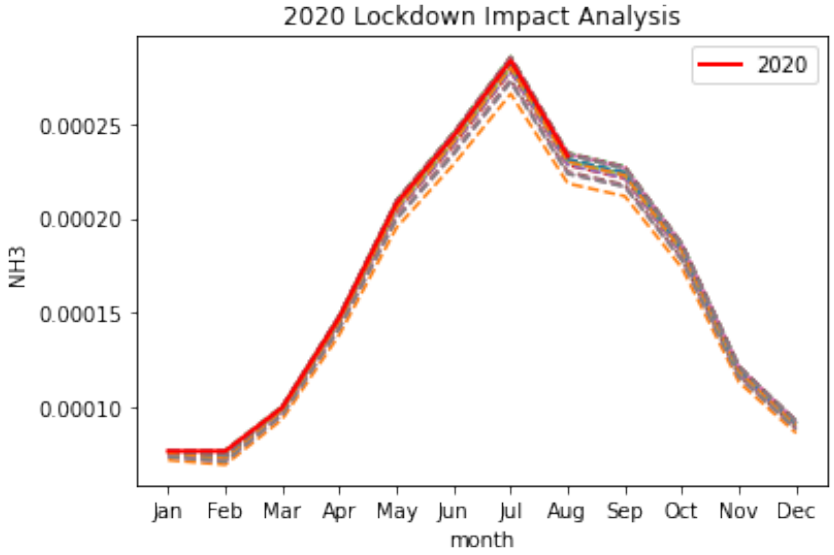

Figura 2.6: Impatto del lockdown sulle emissioni

Terminata la parte di comprensione qualitativa della serie è stata eseguita un'analisi più sistematica di un punto specifico (il punto di massime emissioni), la quale comprende test di normalità, stazionarietà e autocorrelazione, oltre all'applicazione di varie trasformate normalizzanti sui dati. Il codice della seguente sezione è visualizzabile nel file `/notebooks/nh3-ts-analysis.ipynb`.

Lo studio proseguirà con l'analisi della seguente serie storica nel punto di massima emissione.

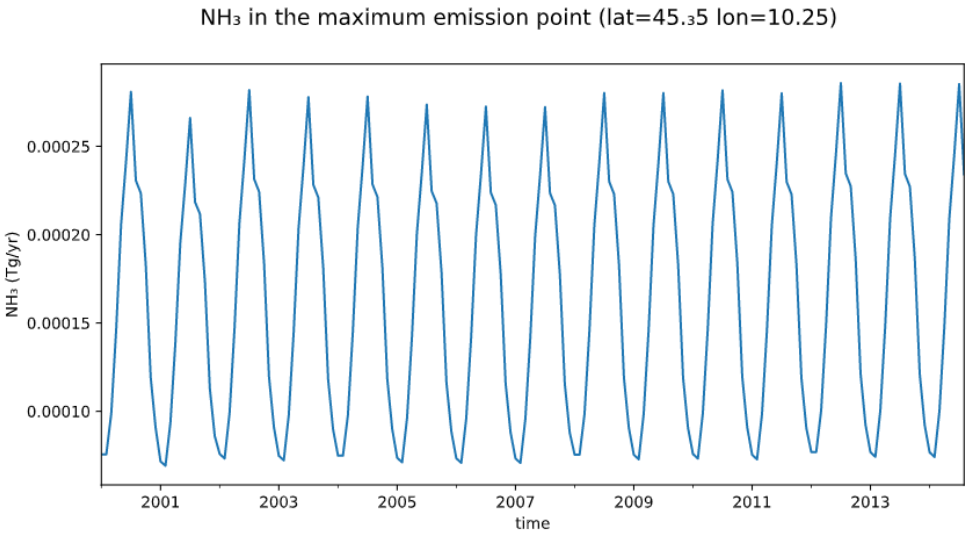

Figura 2.7: Serie storica nel punto di massima emissione

2.2 Normalità

Come primo punto viene testata la **normalità dei dati**. Per fare ciò, è stata dapprima visualizzata la distribuzione dei dati in sovrapposizione a una normale con stessa media e varianza, mostrata nel grafico di sinistra in Figura 2.8. Il grafico di destra rappresenta invece la *Kernel Density Estimation* dell'istogramma, offerta dalla libreria *pandas*: essa è una metodologia non-parametrica per stimare la funzione di densità di probabilità di una variabile casuale, ottenuta utilizzando kernel gaussiani.

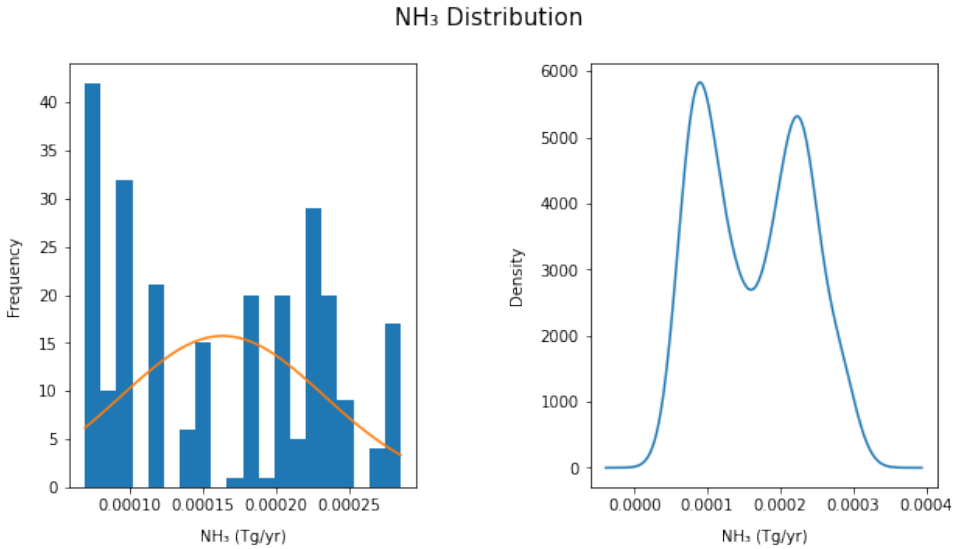

Figura 2.8: Distribuzione dei dati con confronto alla normale

Come si evince dalla figura, i dati non sembrano normali. Questa affermazione viene verificata con due test: il test di *Jarque Bera* e quello di *Lilliefors*.
Jarque Bera utilizza simmetria e curtosi come parametri, e ha come ipotesi nulla che i dati siano normali; la distribuzione è bimodale, incompatibile con la gaussiana in cui l'unica moda corrisponde alla media. Jarque Bera conferma ciò rifiutando l'ipotesi nulla, con un p-value pari a circa $1,99 \cdot 10^{-5}$.
Il test di Lilliefors utilizza la funzione di probabilità cumulata, e anch'esso ha come ipotesi nulla che i dati siano normali. Il test da come risultato un p-value pari a circa $1 \cdot 10^{-3}$ e perciò anch'esso rifiuta l'ipotesi nulla.

È stata quindi poi tentata una **normalizzazione dei dati** della serie, applicando diverse trasformate normalizzanti e rieseguendo i test precedenti ad ogni iterazione, ma nessuna di esse ha portato all'accettazione delle ipotesi nulle. Alcune delle trasformate utilizzate includono: logaritmo, Box Cox, radice quadrata, differenze prime e differenze dodicesime. Quest'ultima è stata eseguita in modo da cercare di eliminare la stagionalità annuale dei dati, ma siccome essi sono quasi deterministici la quasi totalità degli errori convergono a zero, risultando in un picco molto alto, evidente in Figura 2.12.

Le visualizzazioni delle trasformate e i relativi p-value dei test sono riportati nelle figure 2.9, 2.10, 2.11 e 2.12.

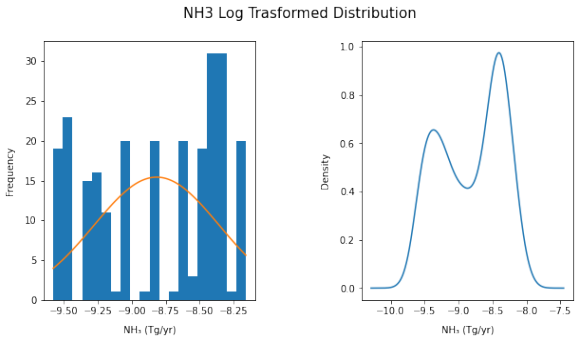

Figura 2.9: Trasformata logaritmo
p-value: $4,35 \cdot 10^{-6}$ (JB), $1 \cdot 10^{-3}$ (LF)

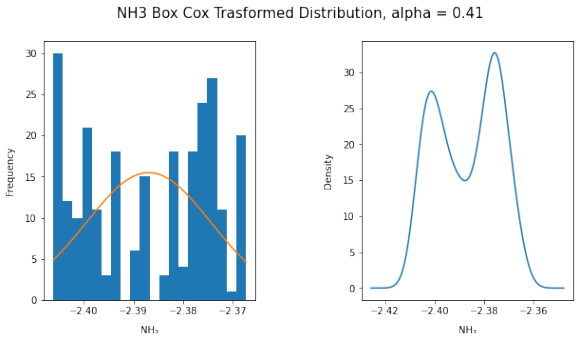

Figura 2.10: Trasformata Box Cox
p-value: $7,52 \cdot 10^{-6}$ (JB), $1 \cdot 10^{-3}$ (LF)

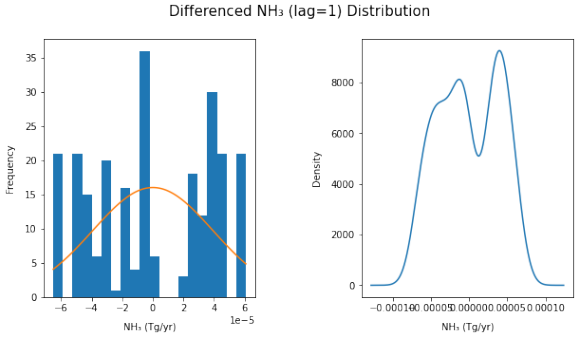

Figura 2.11: Trasformata differenze prime
p-value: $1,18 \cdot 10^{-4}$ (JB), $1 \cdot 10^{-3}$ (LF)

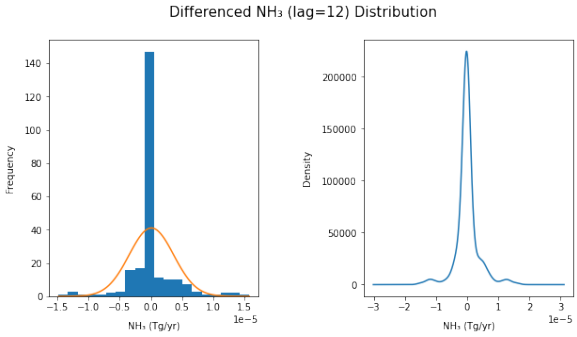

Figura 2.12: Trasformata differenze lag 12
p-value: $4,28 \cdot 10^{-81}$ (JB), $1 \cdot 10^{-3}$ (LF)

Si noti come il test di Lilliefors dà lo stesso p-value ($1 \cdot 10^{-3}$) per tutte le trasformate: questo è dovuto al fatto che esso utilizza dati tabulati che quindi presentano un lower bound, per cui se la statistica scende sotto una certa soglia verrà sempre ritornato un p-value di $1 \cdot 10^{-3}$.

2.3 Stazionarietà

La seconda analisi che si è fatta è stata la **stazionarietà dei dati**, presupposto fondamentale per l'utilizzo di molte procedure statistiche nelle serie storiche. Per la verifica è stato utilizzato il test *Augmented Dickey-Fuller*, il quale ha come ipotesi nulla la presenza di radici unitarie nella serie, con conseguente non stazionarietà dei dati. Anche se visivamente si potrebbe pensare che la serie sia stazionaria, data l'assenza visibile di trend e la regolarità dei dati, il test di Dickey-Fuller accetta l'ipotesi nulla, con un p-value pari a circa $0,3274$. Questo risultato è dovuto alla forte stagionalità delle emissioni, di cui il test non tiene conto. Per ovviare a questo fatto è stata eseguita una destagionalizzazione

della serie (tramite decomposizione, spiegato in sezione 2.5), e rieseguito il test, il quale risponde con rifiuto dell'ipotesi nulla e conseguente stazionarietà, grazie a un p-value pari a circa $4,17 \cdot 10^{-12}$. Questo è importante perché permette di utilizzare modelli più semplici della classe ARIMA con $d = 0$, senza dover lavorare sulle differenze.

2.4 Autocorrelazione

Il terzo punto si concentra sull'**autocorrelazione** della serie storica. Sono stati graficati gli andamenti della funzione di autocorrelazione e della funzione di autocorrelazione parziale, mostrati in Figura 2.13.

Osservando il primo grafico, è evidente come ci siano forti dipendenze stagionali nell'autocorrelazione, le quali seguono l'andamento annuale delle emissioni. Si vede anche che fino a lag 24 l'autocorrelazione rimane estremamente alta, ancora una volta a testimonianza della sinteticità dei dati. Nel terzo grafico si evidenzia come anche l'autocorrelazione a lungo termine sia elevata: anche oltre lag 120, quindi a oltre 10 anni di differenza di serie storica, l'autocorrelazione rimane significativa. Questo risultato è molto importante in quanto si vedrà successivamente che parte di questa autocorrelazione quasi deterministica andrà a trasferirsi inevitabilmente nei residui.

Nel grafico dell'autocorrelazione parziale invece si può vedere una forte dipendenza nei primi due mesi di lag, oltre che a lag 12 e 13 (anch'essi riconducibili alla stagionalità della serie).

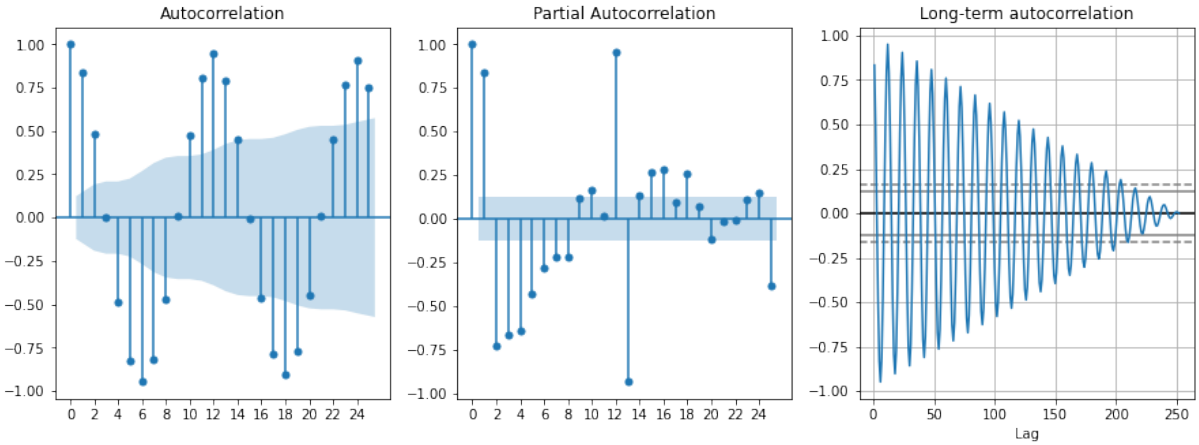

Figura 2.13: Autocorrelazione breve, Autocorrelazione parziale, Autocorrelazione a lungo termine

Sono stati anche generati i *Lag Plots* dei dati da lag 1 a lag 12, i quali come previsto sono risultati anomali. Come mostrato in Figura 2.14, invece della classica "nuvola" di punti che assume una direzione generale a seconda del grado e verso della correlazione, nel caso in esame si hanno una serie di cluster (12 per l'esattezza) contenenti punti estremamente

ravvicinati: questa è una testimonianza della forte correlazione; ogni cluster rappresenta un mese, e ogni anno la correlazione in quel mese è talmente forte che i punti non sono distribuiti in un intorno, ma sono quasi sovrapposti, rappresentando una relazione biunivoca in ogni gruppo.

Analizzando i vari lag plot si osserva una forte correlazione positiva a lag 1 e 2, che va diminuendo verso il lag 3 e 4 (evidenziato dalla forma a ellissoide che si avvicina a una forma circolare). Nei lag 5, 6 e 7 si osserva correlazione negativa (rappresentata bene nella retta che viene a formarsi a lag 6), per poi tornare incorrelata e poi positiva. Si osserva come a lag 12 i dati si dispongano addirittura perfettamente su una retta, risultando in una correlazione pari a 1.

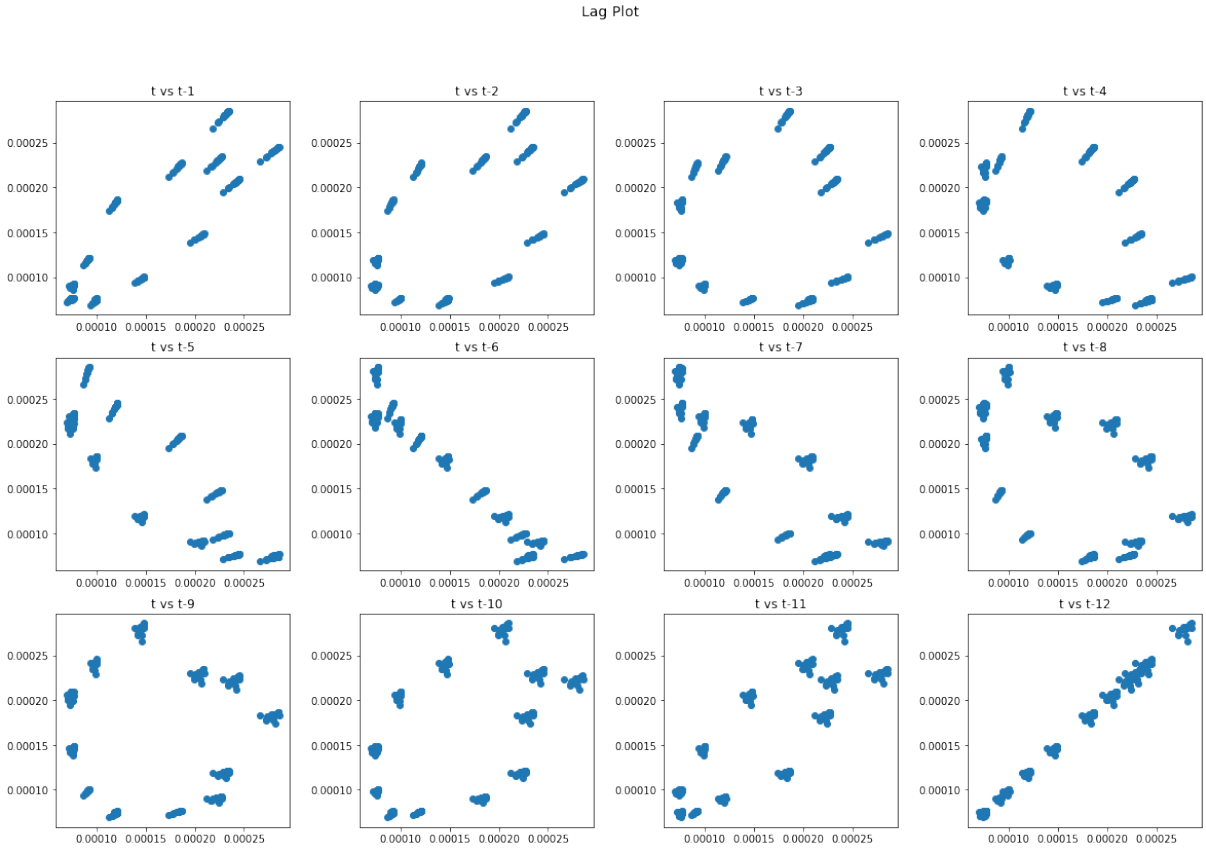

Figura 2.14: Lag Plots della serie storica

2.5 Decomposizione

Come quarta questione si osserva la decomposizione della serie storica per separare additivamente le parti di trend e stagionalità dai dati, evidenziando la parte di residui. Ciò viene effettuato tramite la funzione `seasonal_decompose` di `statsmodels`, i cui risultati sono mostrati in Figura 2.15. Come si osserva, la quasi totalità della decomposizione è rappresentata dalla stagionalità, mentre il trend risulta di difficile interpretazione, essendo instabile ma con valori nell'ordine di $10^{-4} Tg/yr$ in tutta la serie. Anche i residui sono esigui (nell'ordine di $10^{-6} Tg/yr$ in tutta la serie), fatto dovuto all'origine "sintetica" dei dati, ottenuti da un modello preesistente.

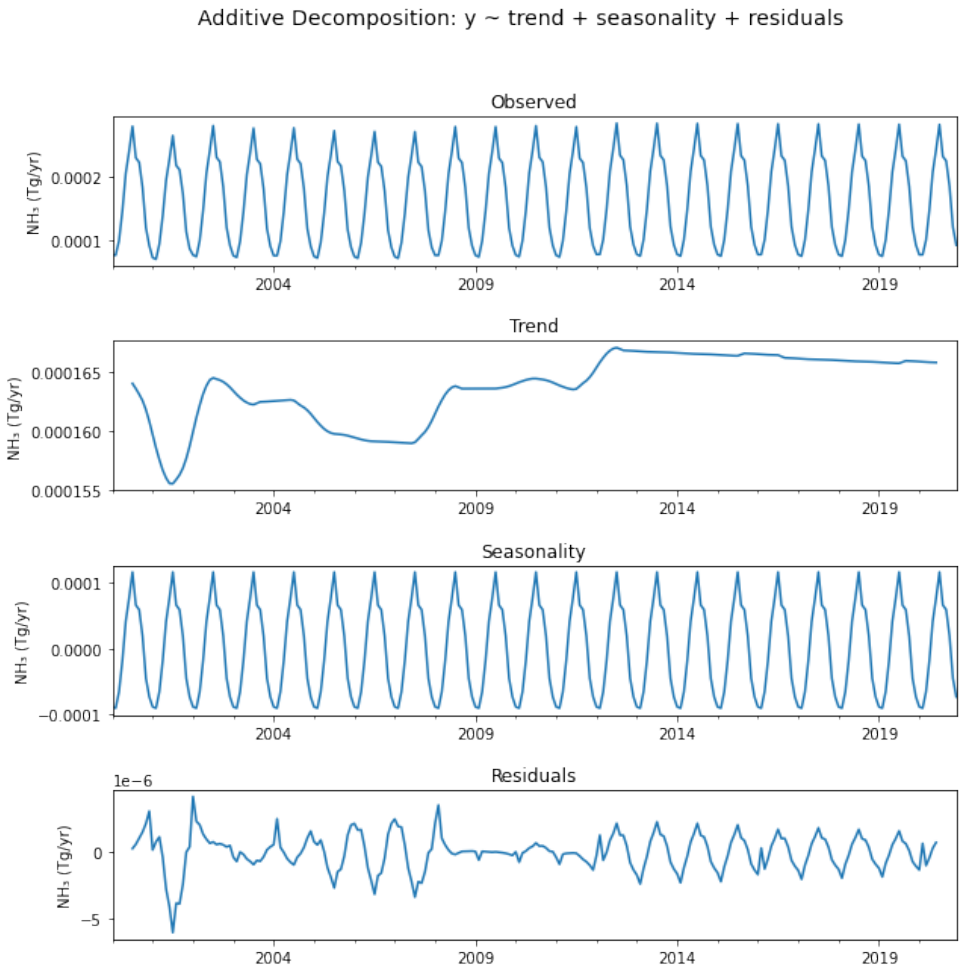

Figura 2.15: Decomposizione additiva della serie storica in trend, stagionalità e residui

Il trend è osservabile più chiaramente nel *Box Plot*, evidenziato in Figura 2.16, che mostra la variabilità annua. Per ogni anno i valori minimo e massimo sono circa gli stessi: le barre del box plot presentano la stessa estensione, a rappresentazione del fatto che l'escursione di emissioni di ammoniaca è simile in tutti gli anni, non presentando una variazione significativa di anno in anno. La media, rappresentata da segmenti di colore verde, mostra come i primi anni i valori oscillino, a testimonianza del fatto che i primi

12 anni di dati sono osservati da dati realmente campionati; oltre questa soglia, invece, le medie diventano estremamente uguali da un anno all'altro.

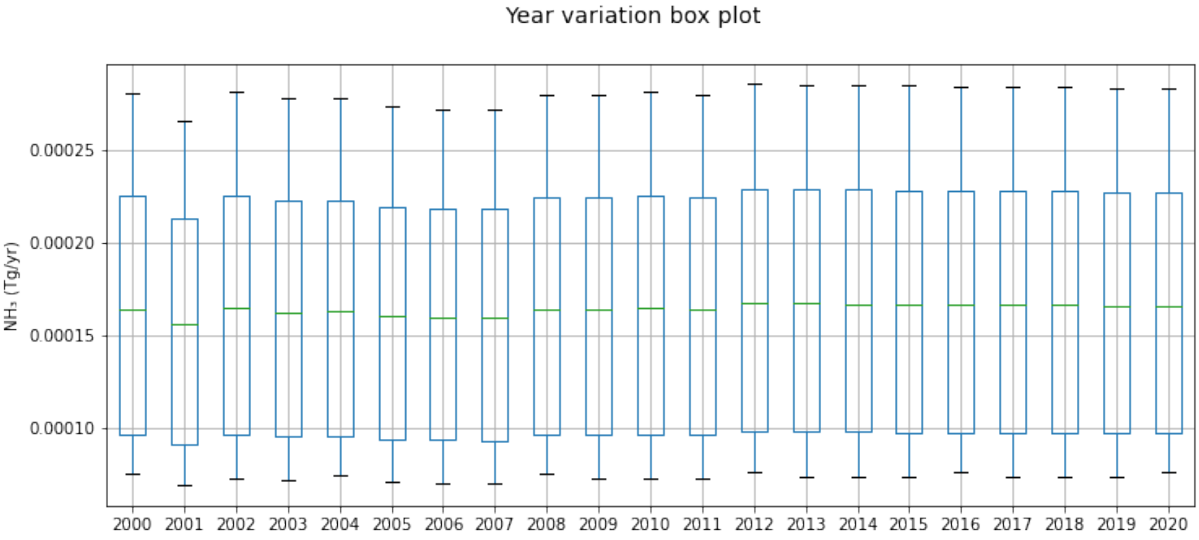

Figura 2.16: Box plot della serie

Capitolo 3

Selezione del modello

In fase di selezione del modello sono stati applicati due metodi differenti con lo scopo di confrontarne i risultati.
Si considera un modello generico SARIMA che, nel corso dell'elaborato, verrà indicato con notazione

$$\text{SARIMA}(p, d, q, P, D, Q)$$

nonché notazione adottata dalla libreria python `statsmodels` per la generazione del modello.
In particolare, i parametri p, d, q fanno riferimento ai coefficienti di un modello ARIMA(p, d, q), mentre P, D e Q modellano la componente stagionale presente a lag s. In questo studio, come visto nei capitoli precedenti, la componente stagionale risulta essere di tipo annuale, quindi $s = 12$.

In entrambi gli approcci adottati spiegati nei prossimi paragrafi, si suddivide il dataset in tre porzioni: un *dataset di training* utilizzato per l'apprendimento del modello, un *dataset di validazione* utilizzato per la valutazione delle performance di tale modello ed un *dataset di test* per verificare le performance del modello su dati non osservati in fase di apprendimento.
I due metodi differiranno principalmente sul criterio di suddivisione del dataset nelle due porzioni di training e validazione.

Entrambi i metodi adottano un procedimento comune: si considerano le possibili combinazioni di parametri caratteristici del modello SARIMAX e, per ogni combinazione di parametri, si stima un modello sulla base del training dataset, si valutano le diverse misure di performance utilizzando, dove necessario, il dataset di validazione e si salvano i risultati.
Dopo aver valutato tutte le possibili combinazioni si analizzano i risultati e si individua il modello che soddisfa al meglio le qualità desiderate.

3.1 Predizione in-sample e out-of-sample

Alcuni criteri di scelta che verranno utilizzati adottano misure ottenute da **predizioni in-sample**, utilizzando il modello per prevedere gli stessi dati utilizzati per apprenderne i parametri, e **predizioni out-of-sample**, utili a prevedere dati non ancora osservati dal modello.
In particolare, le classi di modelli SARIMA della libreria python adottata permettono di effettuare predizione in-sample tramite il metodo `predict`, specificando data di inizio e fine, e di effettuare la predizione out-of-sample dell'istante temporale successivo all'ultimo dato di training con il metodo `forecast`.
In fase di valutazione delle performance, risulta tuttavia interessante avere predizioni out-of-sample di un periodo temporale più esteso. Per fare questo, si è adottato un metodo con re-inserimento delle predizioni effettuate: dopo aver predetto l'istante temporale successivo all'ultimo osservato, si inserisce tale valore all'interno dell'oggetto contenente il modello e lo si considera come valore osservato.
Nel caso si voglia predire dodici valori mensili di emissioni out-of-sample, è quindi possibile effettuare dodici predizioni re-inserendole man mano nel modello. In tal senso, viene definita da noi la funzione `multiple_forecasts`.

3.2 Random Cross Validation

Il primo metodo, in questo elaborato chiamato **Random Cross Validation**, consiste nel campionamento casuale e ripetitivo del dataset di training e di validazione.
In particolare, si campionano casualmente diversi istanti temporali di inizio dell'intervallo di training e, definita a priori l'ampiezza dello stesso (*sliding window*), di definisce l'istante di inizio del dataset di validazione, nonché quello immediatamente successivo all'ultimo istante del dataset di training.
L'aspettativa è quella di ottenere una migliore capacità di generalizzazione del modello.
Per ogni combinazione dei parametri del modello, vengono eseguiti diversi fit del modello, con conseguente rilevazione di diverse misure di prestazione, una per ogni modello.
Ne segue che per ogni combinazione di parametri si analizzerà il valor medio delle misure di prestazione ricavate dalle diverse iterazioni di campionamento casuale e fitting del modello.

Nella figura Figura 3.1 si è rappresentato il funzionamento.

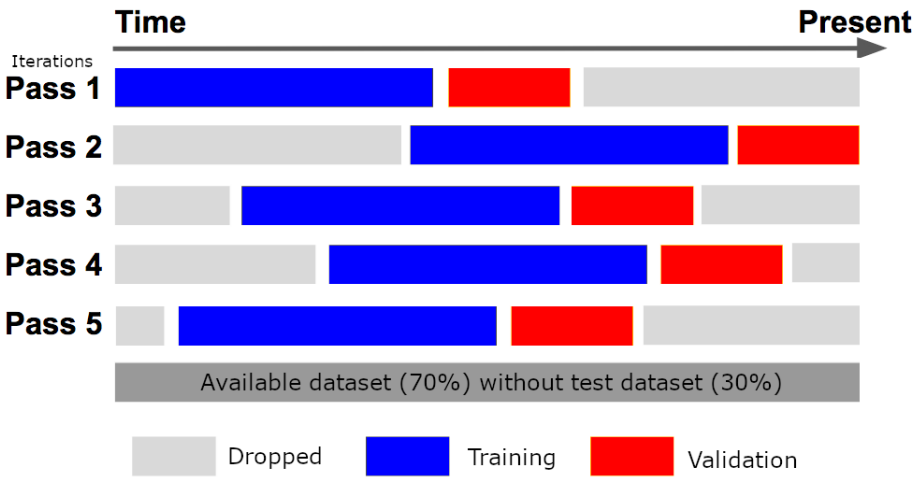

Figura 3.1: Schema della Random Cross Validation

Si è quindi definita la funzione `random_cross_validation`, in cui è possibile specificare il numero di iterazioni da svolgere per ogni combinazione di parametri. Naturalmente, nei limiti delle capacità e di tempo computazionale, tanto maggiore è il numero di iterazioni, tanto maggiore sarà l'effetto di generalizzazione desiderato.
In questo caso, l'ampiezza della finestra di valori per il dataset di apprendimento è stata definita pari a 10 anni, mentre quella di validazione pari a 2, in modo da considerare e valutare la capacità del modello di catturare un'intera stagionalità.

3.3 Rolling Cross Validation

Quando i dati non sono indipendenti tra loro (come nel caso di serie storiche), la cross-svalidazione diventa più difficile poiché tralasciare un'osservazione non rimuove tutte le informazioni ad essa associate a causa delle correlazioni con altre osservazioni.
Per la previsione delle serie storiche può essere usata la **Rolling Cross Validation**, descritta in un articolo di Rob J Hyndman[1].

Il funzionamento generale di questa procedura è il seguente: si vuole avere una serie di test sui cui fare la cross validazione: per ogni test di questa serie si ha un *validation set* costituito da una singola osservazione e il *training set* formato dalle osservazioni che precedono temporalmente l'osservazione contenuta nel set di validazione.
Pertanto, nessuna osservazione futura potrà essere utilizzata per costruire la previsione.
Nella figura Figura 3.2 sono presenti una serie di test: i rispettivi training sets sono colorati in blu, mentre i sets di validazione sono colorati in rosso.

[1]Rob J Hyndman, *"Cross-validation for time series"* (https://robjhyndman.com/hyndsight/tscv/)

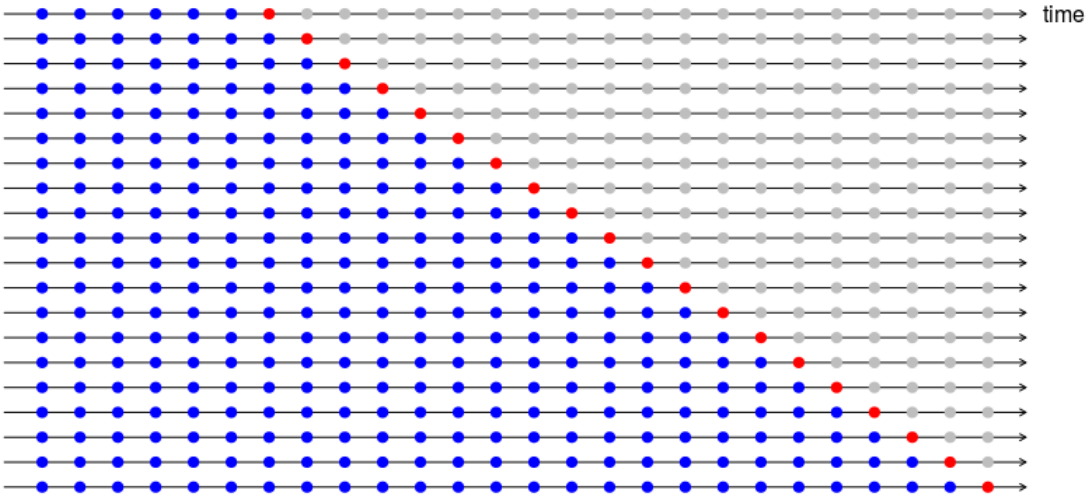

Figura 3.2: Schema della Rolling Cross Validation

L'accuratezza della previsione viene verificata calcolando la media degli indici di perfomance su tutti i gruppi di test.

In questo caso si è dovuto creare la funzione `rolling_cross_validation` che riceve come input la serie storica delle emissioni di ammoniaca a partire dal gennaio 2000 fino a agosto 2014 (il 70% della serie storica viene usato come training): i restanti anni sono stati tenuti per effettuare la validazione del modello. Questa funzione verrà chiamata in ogni iterazione, per ogni modello SARIMA usato nel model selection.

Nella prima iterazione si parte con i primi 36 mesi della serie storica come training set per non avere troppo pochi dati di training iniziali.
Inoltre, a differenza del caso generale, il validation set è costituito non da una sola osservazione ma da 12 osservazioni. Si vuole quindi effettuare la previsione dell'anno successivo ma non è possibile effettuare tutte le 12 previsioni mensili in un colpo solo: bisogna quindi ricorrere alla funzione `multiple_forecasts`, descritta in precedenza, per fare predizioni out-of-sample annuali re-inserendo di volta in volta le predizioni mensili effettuate.

Quindi, a differenza della Random Cross Validation, anziché campionare sempre lo stesso intervallo temporale tante volte modificando il mese d'inizio, si incrementa il training set ad ogni iterazione e si va a effettuare la predizione out-of-sample dell'anno successivo.

3.4 Criteri di scelta e misure di performance

Per poter valutare la bontà del modello sono stati considerati quattro differenti criteri: **AIC**, **BIC**, **Root Mean Squared Error** delle previsioni out-of-sample sul da-

taset di validazione e **Mean Absolute Autocorrelation** sui residui delle predizioni in-sample.

Mentre i primi tre forniscono informazioni relative alla bontà previsionale del modello, eventualmente penalizzandone la numerosità dei parametri, l'ultimo permette di ottimizzare il modello in funzione delle correlazioni presenti nei residui.
Nel caso in esame, infatti, la forte correlazione presente nei dati si trasporta anche nei residui del modello. Il modello interpola perfettamente i dati, comportamento prevedibile ed atteso dal momento che sono deterministicamente ottenuti da un ulteriore modello generatore.
Tuttavia, vengono compiuti piccolissimi errori periodicamente in prossimità dei picchi, tradotti in stagionalità e correlazione nei residui di previsione. Osservando la correlazione media (ed in valore assoluto) delle correlazioni degli errori nei primi 24 istanti temporali, è possibile individuare il modello che lascia meno correlazione nei residui.

In funzione dei risultati, possono essere individuati quattro diversi modelli, candidati ad essere i migliori, uno per ogni criterio scelto. Talvolta può capitare che più criteri siano di comune accordo sulla scelta del modello migliore, come spesso accade con i modelli proposti per minor AIC e minor BIC.

Considerando come misura di prestazione principale l'AIC, sono stati inoltre considerati i modelli con un AIC nell'intorno del minimo trovato. Di questi si possono osservare le altre misure di prestazione e valutare eventuali trade-off, ossia di perdita di prestazione secondo AIC a favore di un minore errore di validazione o di una minore autocorrelazione media negli errori di previsione dei dataset di validazione usati, consapevoli del fatto che si stanno comunque osservando i modelli a minor AIC tra tutti quelli testati.

Capitolo 4

Analisi dei risultati

In questo capitolo vengono riportati i risultati del processo di selezione del modello nelle due varianti proposte.
In entrambi i metodi, il dataset di test equivale al 30% della serie storica, ed il 70% rimanente viene suddiviso tra training e validazione.
I parametri p, q, P, Q sono selezionati dai seguenti possibili set.

	p	q	P	Q
Set	(1,2,3,4)	(0,1,2)	(0,1)	(0,1)

Da notare che includendo la possibilità che il parametro P assuma valore 0, si valuta anche il caso in cui non sia presente stagionalità.

Una volta selezionata la struttura del modello, entrambi i metodi prevedono l'apprendimento del modello individuato su tutto il dataset di training (nonchè 70% della serie storica).

4.1 Risultati ottenuti con random cross validation

I risultati del processo di random cross validation, in seguito riportati, sono ottenuti mediante selezione casuale di 20 istanti temporali, come illustrato nel capitolo 3.2. A partire da ogni istante temporale individuato, i successivi 120 mesi costituiscono il dataset di training ed i seguenti 24 quello di validazione. La scelta di utilizzare due anni per la validazione è motivata dall'interesse di valutare come venga incorporata nel modello la stagionalità: essendo questa annuale, è opportuno validare almeno un'intera stagionalità.

In Figura 4.1 si riportano i risultati di AIC, BIC, RMSE e autocorrelazione media. Come si può osservare, i risultati di AIC, BIC sono molto simili tra loro, mentre RMSE e autocorrelazione media favoriscono modelli con un ordine autoregressivo maggiore.

I modelli migliori per misura di prestazione sono riportati nella tabella sottostante

	AIC	BIC	RMSE	MAC
Order	(2,0,0,1,0,0,12)	(2,0,0,1,0,0,12)	(4,0,0,1,0,0,12)	(4,0,0,1,0,0,12)

Come atteso, il modello proposto da ogni misura di prestazione propone di inserire una stagionalità annuale.

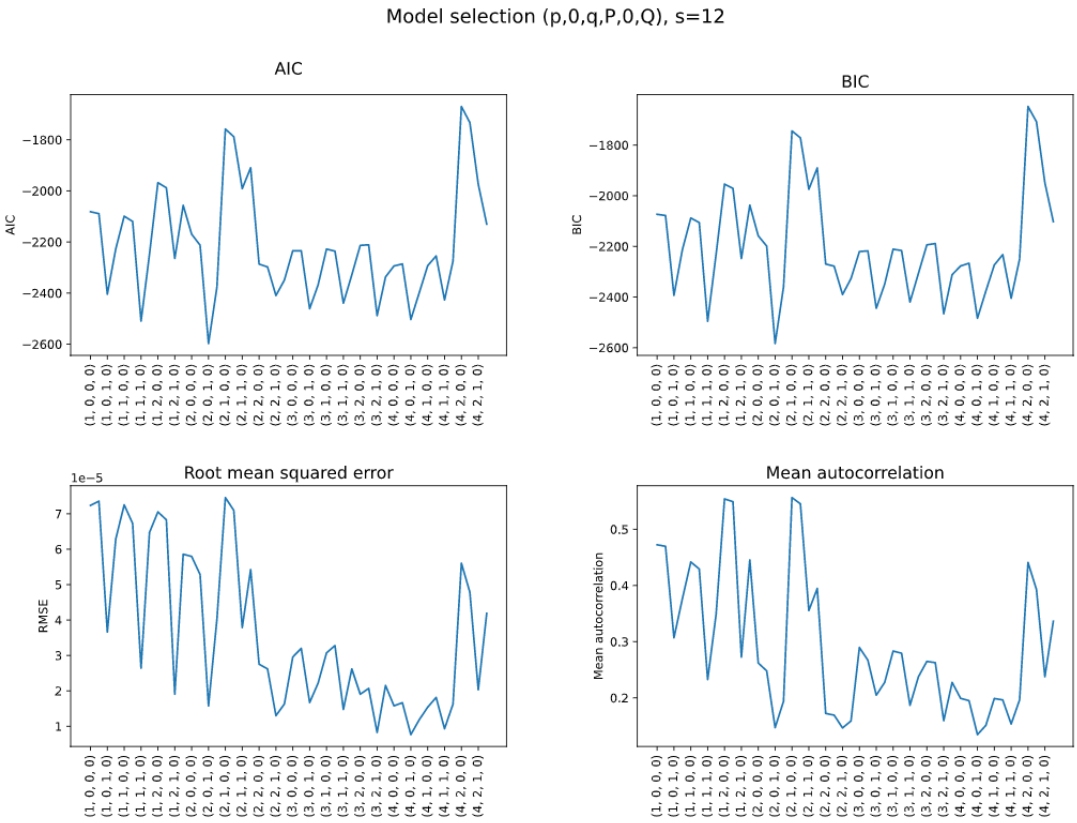

Figura 4.1: AIC, BIC, RMSE e autocorrelazione media per ogni combinazione di parametri testata

A questo punto, può essere di particolare interesse osservare le prestazioni dei modelli nell'intorno del minimo AIC. In questo elaborato si è sempre assunto AIC (e similmente BIC) come misura di prestazione principale. Come riportato in Figura 4.2, osservando nell'intorno del minimo AIC si può notare la presenza della soluzione proposta dalla misura di prestazione RMSE, e che l'RMSE della soluzione con miglior AIC si discosta di circa $5 \cdot 10^{-6}$ dal minimo RMSE.

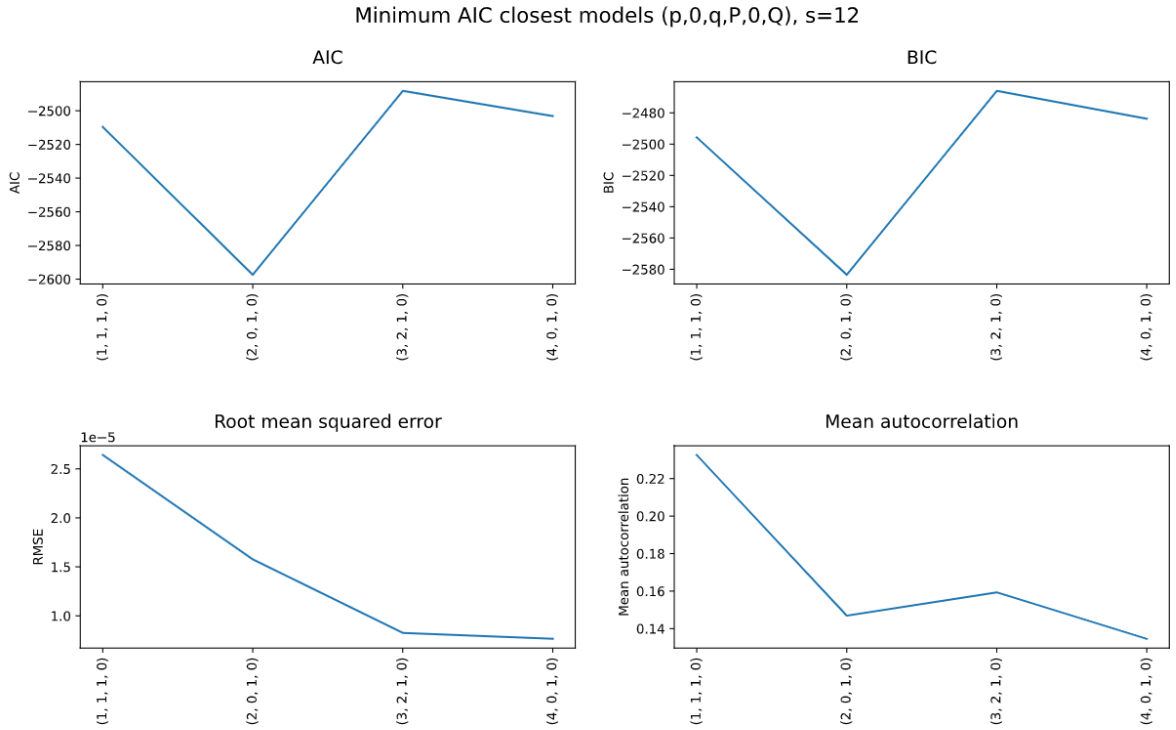

Figura 4.2: AIC, BIC, RMSE e autocorrelazione media nell'intorno del minimo AIC

Il miglior modello selezionato risulta essere:

$$\text{SARIMA}(2, 0, 0, 1, 0, 0, 12)$$

	const	ar.L1	ar.L2	ar.S.L12	σ^2
value	0.000163	1.453433	-0.737639	0.999624	2.838353e-12
pvalue	0.224917	0.000000	0.000000	0.000000	9.847158e-01

$$y_t = 1.63 \cdot 10^{-4} + 1.45 y_{t-1} + -0.74 y_{t-2} + y_{t-12} + \epsilon_t \tag{4.1}$$

4.2 Risultati ottenuti con rolling cross validation

I risultati del processo di rolling cross validation in seguito riportati sono ottenuti mediante selezione successiva di istanti temporali incrementali, come illustrato nel capitolo 3.3. La lunghezza dell'intervallo temporale di apprendimento iniziale è di 120 mesi, i cui seguenti 24 costituiscono quello di validazione. Ad ogni iterazione viene incrementata l'ampiezza del dataset di training di 12 mesi, fino ad esaurimento di dati disponibili. La scelta di utilizzare due anni per la validazione è motivata dall'interesse di valutare come venga

incorporata nel modello la stagionalità: essendo questa annuale, è opportuno validare almeno un'intera stagionalità.

Come nel caso precedente, dalla Figura 4.3 si evincono risultati identici per AIC e BIC, che risultano essere condivisi dai modelli migliori per RMSE, mentre MAC propone modelli con un numero di parametri superiore.
I modelli migliori per misura di prestazione sono riportati nella tabella sottostante.

	AIC	BIC	RMSE	MAC
Order	(2,0,0,1,0,0,12)	(2,0,0,1,0,0,12)	(2,0,0,1,0,0,12)	(4,0,0,1,0,0,12)

Come atteso, il modello proposto da ogni misura di prestazione propone di inserire una stagionalità annuale.

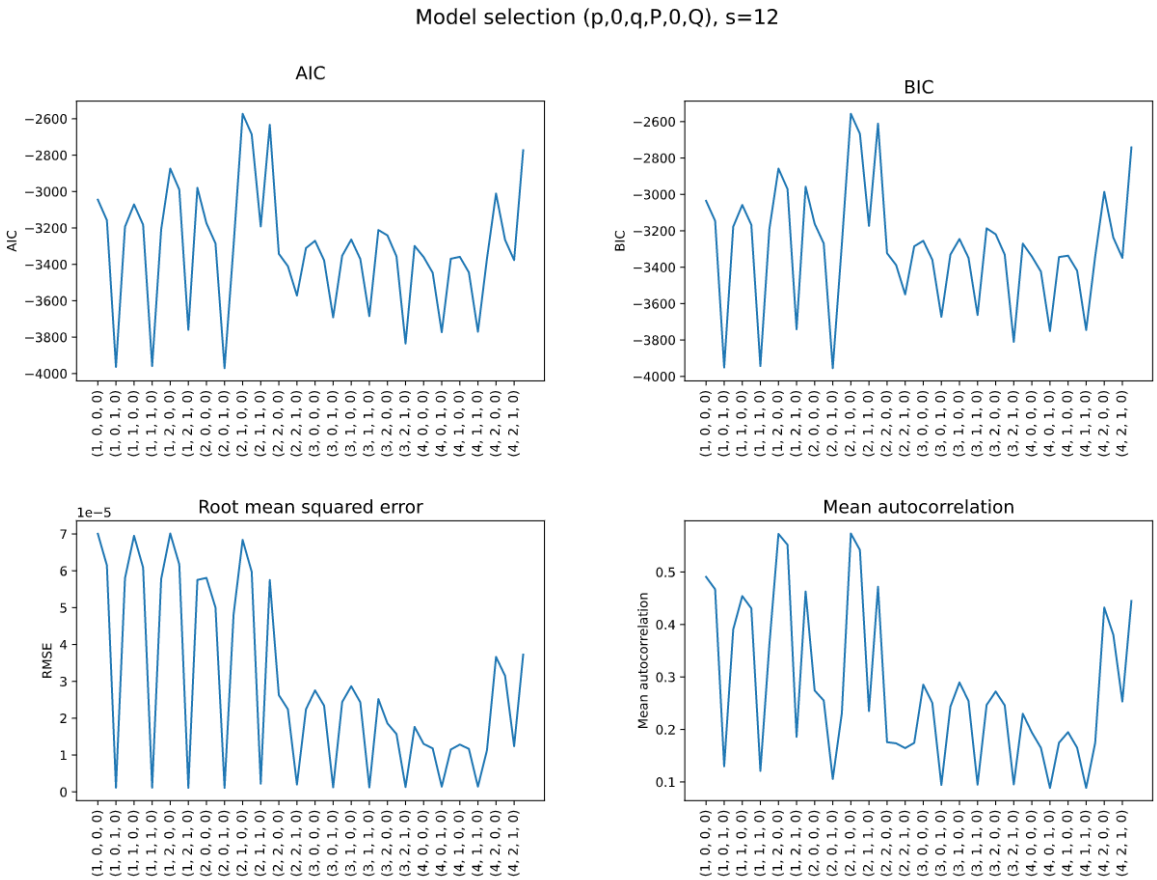

Figura 4.3: AIC, BIC, RMSE e autocorrelazione media per ogni combinazione di parametri testata

Esplorando nell'intorno del minimo AIC (Figura 4.4), si conferma la decisione univoca di AIC, BIC e RMSE. Il miglior modello per AIC ed il miglior modello per MAC risultano coincidenti con quelli individuati dal metodo random cross validation prima

presentato. In questo caso, però, la prestazione del modello è condivisa da più misure di prestazione.

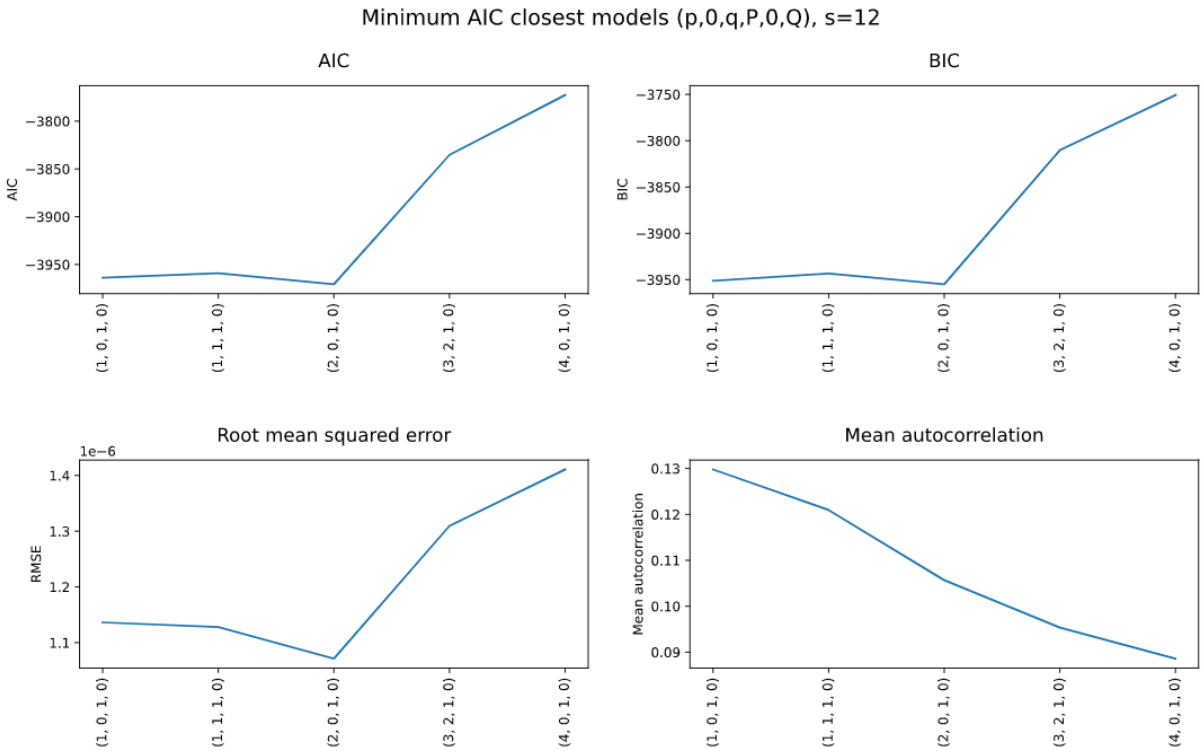

Figura 4.4: AIC, BIC, RMSE e autocorrelazione media nell'intorno del minimo AIC

Essendo il modello individuato analogo al caso precedente, ed essendo uguale il dataset di training finale, il modello selezionato risulta essere:

$$\text{SARIMA}(2, 0, 0, 1, 0, 0, 12)$$

	const	ar.L1	ar.L2	ar.S.L12	σ^2
value	0.000163	1.453433	-0.737639	0.999624	2.838353e-12
pvalue	0.224917	0.000000	0.000000	0.000000	9.847158e-01

$$y_t = 1.63 \cdot 10^{-4} + 1.45y_{t-1} + -0.74y_{t-2} + y_{t-12} + \epsilon_t \tag{4.2}$$

Capitolo 5

Analisi dei residui

Una volta selezionato e fittato il miglior modello per la serie storica in esame, si è proceduto con l'analisi dei residui delle predizioni out-of-sample con reinserimento dei valori sul dataset di test (30%), ignoto e non osservato dal modello precedentemente.
Essendo la serie storica di bassissima variabilità nel tempo, fatto dovuto alla sua natura deterministica, le previsioni risultano quasi perfette.
A causa di questo livello inevitabile di overfitting, i residui non riportano le proprietà di omoschedasticità, normalità e incorrelazione desiderati.

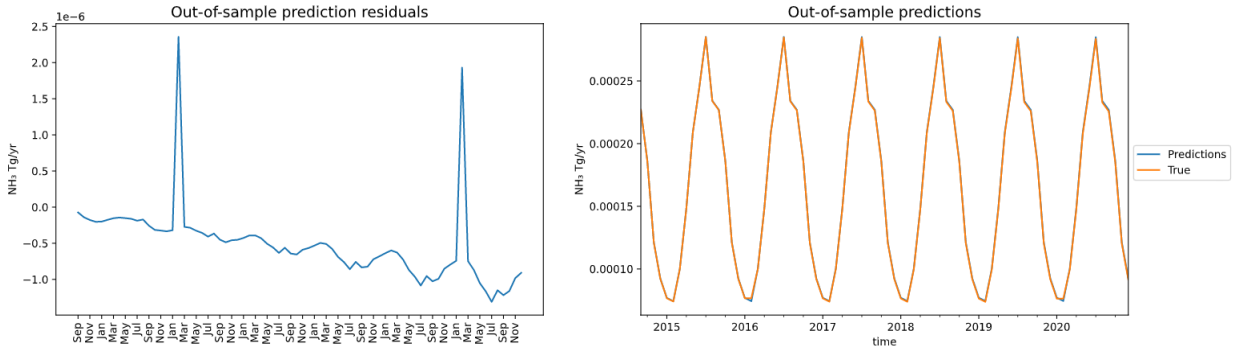

Figura 5.1: Previsioni ed errori di previsione out-of-sample

Le previsioni seguono perfettamente la serie storica vera, con lievi errori nei picchi di minima emissione, in corrispondenza del mese di febbraio. Nella Figura 5.2 si riporta la correlazione dei residui. Specialmente entro i primi 12 lag, la forte correlazione della serie storica si riflette anche nei residui.

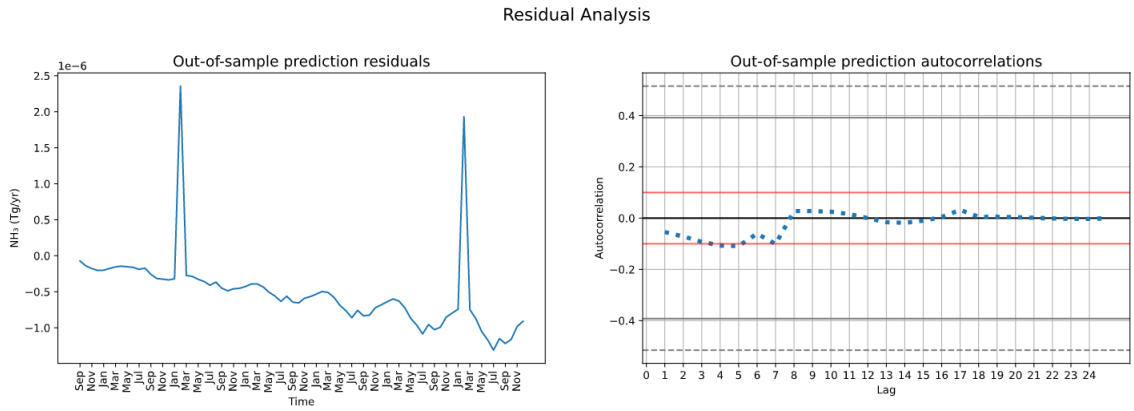

Figura 5.2: Previsioni ed errori di previsione out-of-sample

Infine, i test di omoschedasticità (differenza di varianza tra i primi e gli ultimi istanti temporali), di normalità (sia Jarque-Bera che Lilliefors) e di incorrelazione (Ljungbox con informazione sul numero di parametri stimati) forniscono p-value bassi, con conseguente rigetto dell'ipotesi nulla.

Ad esempio, in Figura 5.3 è riportata la distribuzione degli errori. La perfezione della stima concentra gli errori attorno allo zero, elevando e stringendo la campana e rendendo la distribuzione non comparabile ad una gaussiana.

Per quanto riguarda l'omoschedasticità, si nota come gli errori siano prossimi allo zero con picchi nei mesi di febbraio che ne amplificano la varianza.

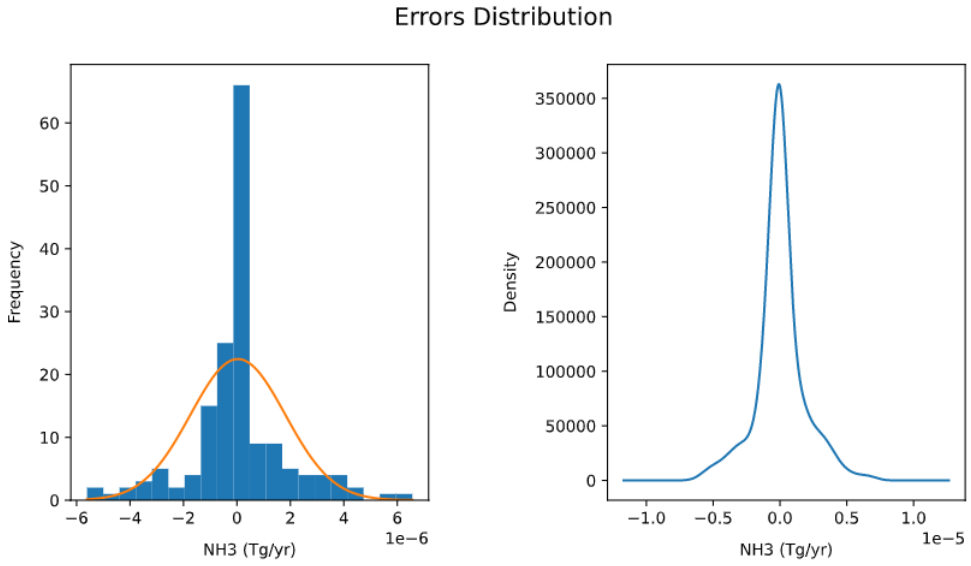

Figura 5.3: Distribuzione degli errori di previsione out-of-sample

Capitolo 6

Regressione con armoniche ed errori SARIMA

È stato quindi effettuato un tentativo di miglioramento del modello previsionale della serie storica utilizzando delle **armoniche** come regressori in modo da catturare la stagionalità e modellando i residui attraverso un modello SARIMA.
Per la selezione del modello, è stato applicato il metodo di random cross validation, cross-validando gli stessi parametri menzionati nel capitolo 3.2 con l'aggiunta dell'ordine delle armoniche da inserire come regressori.
In particolare, essendo la stagionalità a 12 mesi, i possibili ordini cross-validati sono da 1 a 6. Anche in questo caso, la stagionalità dei residui non è imposta e si osserva la sua influenza sulle prestazioni del modello dai risultati ottenuti.

	AIC	BIC	RMSE	MAC
Order	(3,0,0,1,0,1,12)	(3,0,0,1,0,0,12)	(2,0,0,1,0,1,12)	(3,0,2,1,0,0,12)
Harmonics	4	4	1	2

Analogamente a quanto fatto in precedenza, la Figura 6.1 mostra i risultati in termini di misure di performance.

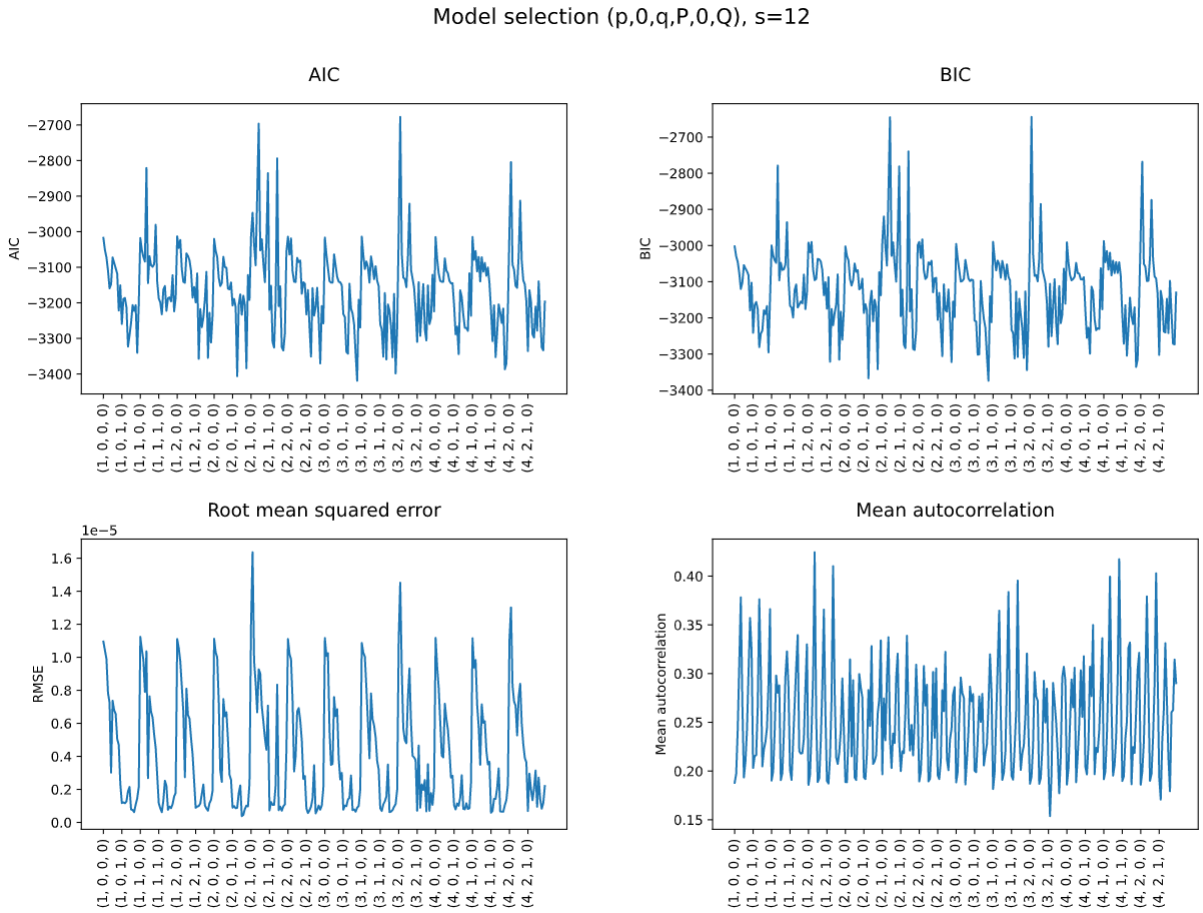

Figura 6.1: AIC, BIC, RMSE e autocorrelazione media per ogni combinazione di parametri testata

Il miglior modello selezionato risulta avere armoniche di ordine 4 ed i residui modellati da:

$$\text{SARIMA}(3, 0, 0, 1, 0, 1, 12)$$

	const	sin(t)	cos(t)	sin(2t)	cos(2t)
value	0.000163	-5.850427e-05	-7.697028e-05	7.449483e-07	0.000005
p-value	0.000018	9.952371e-146	2.306730e-230	6.179992e-01	0.005626

	sin(3t)	cos(3t)	sin(4t)	cos(4t)
value	8.750962e-07	-0.000005	0.000008	-6.701602e-07
p-value	6.039491e-01	0.001596	0.004447	8.094182e-01

	ar.L1	ar.L2	ar.L3	ar.S.L12	ma.S.L12	σ^2
value	-0.662795	0.758343	0.80004	0.965232	-0.340646	5.956550e-12
p-value	0.000000	0.000000	0.00000	0.000000	0.000000	9.647668e-01

$$\begin{aligned} y_t = &+ 1.63 \cdot 10^{-4} - 0.66 y_{t-1} + 0.76 y_{t-2} + 0.80 y_{t-3} + 0.97 y_{t-12} - 0.34 \epsilon_{t-12} + \epsilon_t \\ &- 5.85 \cdot 10^{-5} \sin(t) - 7.7 \cdot 10^{-5} \cos(t) + 7.45 \cdot 10^{-7} \sin(2t) + 5.0 \cdot 10^{-6} \cos(2t) \\ &+ 8.75 \cdot 10^{-7} \sin(3t) - 5.0 \cdot 10^{-6} \cos(3t) + 8.0 \cdot 10^{-6} \sin(4t) - 6.7 \cdot 10^{-7} \cos(4t) \end{aligned}$$

I grafici sotto riportati mostrano come gli errori risultino più distribuiti, nonostante si intravedano comunque picchi nei mesi di febbraio, come nei metodi precedenti. Nonostante le previsioni siano molto accurate, il modello non cattura tanta informazione quanta quella catturata dal modello senza regressori, fatto evidenziato dalle forti correlazioni presenti negli errori di previsione.

Anche in questo caso, gli errori non risultano essere distribuiti normalmente, come dimostrato dai test di Jarque-Bera e Lilliefors utilizzati. Allo stesso modo, non risultano omoschedastici e sono ancora più evidenti e forti le correlazioni.

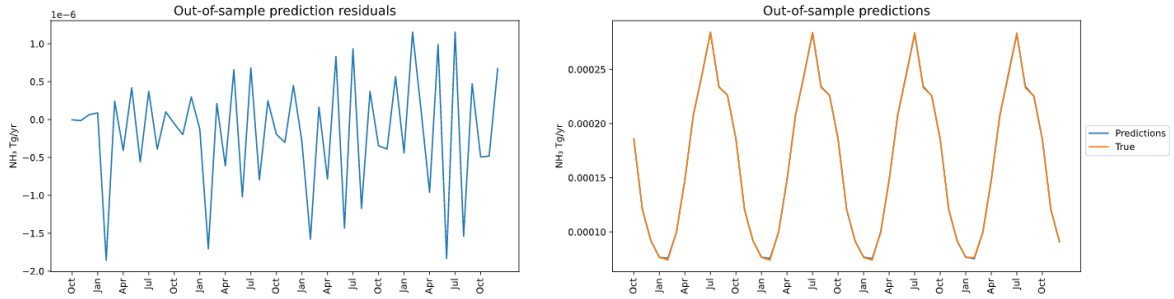

Figura 6.2: Previsioni ed errori di previsione out-of-sample con modello di regressione ad armoniche ed errori SARIMA

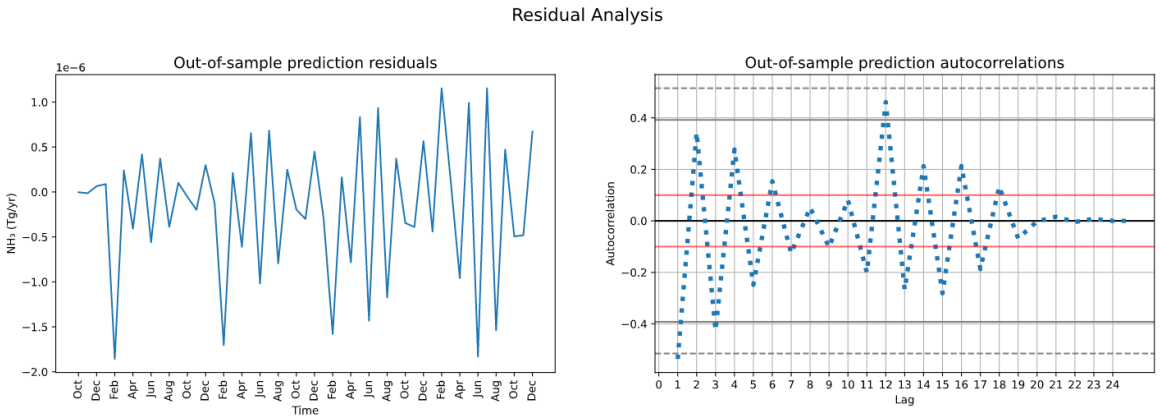

Figura 6.3: Autocorrelazione degli errori di previsione out-of-sample con modello di regressione ad armoniche ed errori SARIMA

Capitolo 7

Applicazione del modello nei punti di minima e media emissione

Si pone ora il problema di capire se il modello individuato a partire dalla serie storica nel punto di massima emissione sia applicabile anche per previsione nei punti di minima e media emissione, come presentato nel capitolo 1.
In entrambi i casi, è stato selezionato il modello migliore individuato con random cross validation (vedi sezione 3.2) ed è stato nuovamente fittato sulla nuova serie storica da analizzare. Questo perchè, esistendo dipendenze temporali, utilizzando un modello già fittato su una scala differente si otterrebbero previsioni out-of-sample appartenenti a tale scala.

Per quanto riguarda il punto di **minima** emissione, la scala della serie storica (in Figura 7.1) risulta di 3 ordini di grandezza inferiore rispetto a quella analizzata fino ad ora.

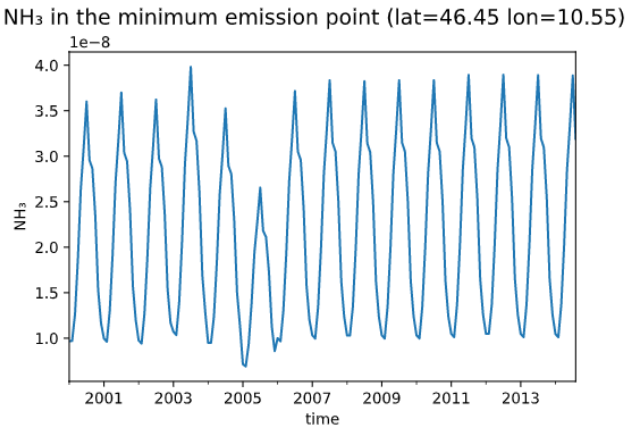

Figura 7.1: Serie storica nel punto di minima emissione

Applicando il modello presentato nell'equazione 4.1, fittandolo con il 70% della serie storica ed effettuando previsioni out-of-sample sul 30% dei dati rimanenti si ottengono i risultati qui riportati.

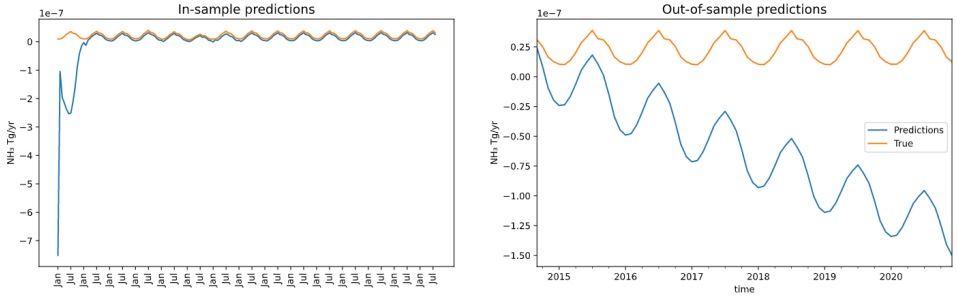

Figura 7.2: Previsioni nel punto di minima emissione

In primo luogo, si nota come le predizioni in-sample del primo anno siano completamente errate, evento causato dalla mancanza di valori passati per poter sfruttare la dipendenza stagionale che, come si è visto nei risultati precedenti, risulta avere un impatto importante. Dopo il primo anno le predizioni si assestano. Tuttavia, le predizioni out-of-sample mostrano risultati scadenti, probabilmente dovuti alla sensibilità della scala, nell'ordine di 10^{-7}. Ciò è dovuto anche al coefficiente della componente autoregressiva della parte stagionale inferiore ad 1 (0.96), causa del continuo smorzamento e sottostima delle previsioni.

Per quanto riguarda la serie storica nel punto di **media** emissione (vedi Figura 7.3), l'ordine di grandezza è analogo a quello del punto di massimo, con valori in media la metà della serie storica su cui si è svolto lo studio.

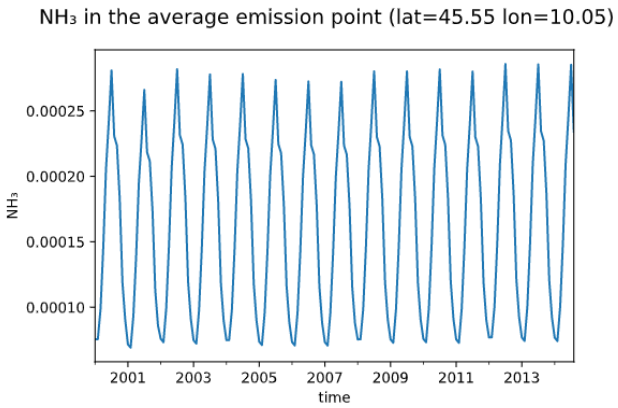

Figura 7.3: Serie storica nel punto di media emissione

Le previsioni sono mostrate nelle figura Figura 7.4. Rispetto al caso del punto di minimo, gli errori di previsione in-sample nel primo anno sono meno rilevanti ma comunque

presenti. Le prestazioni out-of-sample risultano di pari qualità rispetto a quelle viste nel caso di massima emissione dovute anche al fatto che il coefficiente autoregressivo della parte stagionale è esattamente 1.

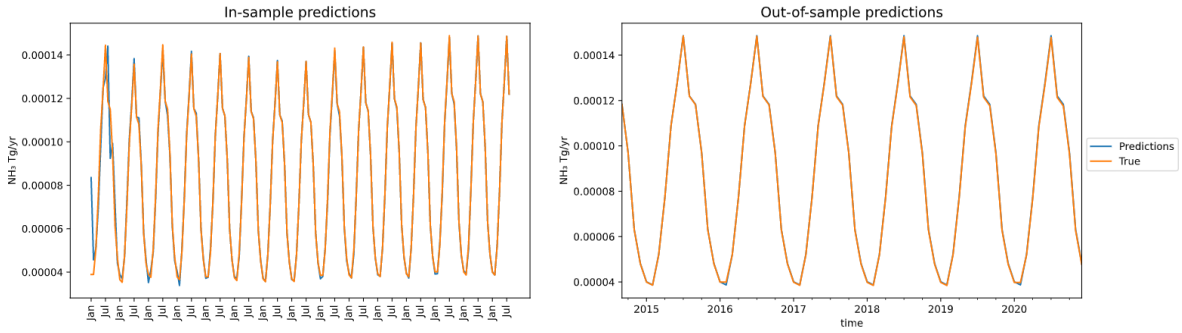

Figura 7.4: Previsioni nel punto di media emissione

Capitolo 8

Sviluppo in Python

8.1 Librerie usate

Illustriamo le principali librerie e pacchetti Python usati nel progetto e alcuni confronti con le alternative in MATLAB[1].

8.1.1 xarray

Il pacchetto *xarray*[2] è un progetto open source utile per lavorare con array multidimensionali in modo semplice ed efficiente.
Grazie a questa libreria si è potuto importare i dati provenienti da files in formato NetCDF o Shapefile.

8.1.2 pandas

Il pacchetto *pandas*[3] è uno strumento open source che permette la manipolazione e l'analisi dei dati in modo veloce, potente e flessibile.
Questa libreria è stata usata per trattare la serie storica dell'ammoniaca e creare il dataframe contenente i regressori (armoniche).

8.1.3 matplotlib

Il pacchetto *matplotlib*[4] è una libreria per creare grafici statici, animati e interattivi in Python.
Questa libreria è stata usata per generare tutti i grafici presenti nel progetto, riuscendo

[1] MATLAB (https://it.mathworks.com/products/matlab.html)
[2] xarray (http://xarray.pydata.org/)
[3] pandas (https://pandas.pydata.org/)
[4] matplotlib (https://matplotlib.org/)

a gestire un'ampia varietà di dati come input e permettendo la personalizzazione di ogni dettaglio del grafico quale colori, legenda, scala, sovrapposizioni di immagini...

8.1.4 numpy

Il pacchetto *numpy*[5] è fondamentale per il calcolo scientifico in Python.
NumPy offre diverse funzioni tra cui: numerose funzioni matematiche di base, generazione casuale di numeri, algoritmi di algebra lineare, trasformate di Fourier...

8.1.5 statsmodels

Il modulo Python *statsmodels*[6] (versione 0.12.1) fornisce una serie di classi e funzioni per la stima di numerosi modelli statistici differenti, sia per condurre tests statistici che data exploration.
Supporta modelli in formato `pandas DataFrame`.

Una delle classi principali usata per creare i modelli è `statsmodels.tsa.arima_model.ARIMA` che gestisce i modelli del tipo SARIMAX$(p, d, q) \times (P, D, Q, s)$, potendo quindi specificare eventuali stagionalità, differenze prime o variabili esogene.
Alcuni metodi usati di questo pacchetto:

- Metodo `ARIMA.fit` della classe `statsmodels.tsa.arima_model.ARIMA` usato per fittare il modello attraverso la massima verosimiglianza basandosi sul filtro di Kalman. Ritorna un oggetto della `statsmodels.tsa.arima_model.ARIMAResults`.
  Il metodo `ARIMA.fit` è equivalente al metodo `estimate` usato in MATLAB per stimare il modello.
  L'algoritmo di stima usato da `ARIMA.fit` calcola i parametri basandosi sulla massima verosomiglianza ricavata attraverso un filtro di Kalman: si riporta la serie storica in un dominio "state space" che permette di calcolare la verosomiglianza in modo più efficente e veloce.
  Tuttavia il filtro di Kalman ha una grossa dipendenza dai dati iniziali: probabilmente per questo motivo si sono riscontrati risultati molto strani come descritto nel capitolo 4; un modo per ovviare a questi problemi è usare serie più lunghe in modo da eliminare gli effetti dei dati iniziali e poter convergere ad uno stato stazionario (nel nostro caso serve circa 5 anni per arrivare a convergenza).
- Metodo `ARIMAResults.forecast` della classe `statsmodels.tsa.arima_model.ARIMAResults` che permette di fare la previsione out-of-sample.

[5] numpy (https://numpy.org/)
[6] statsmodels (https://www.statsmodels.org/)

- Metodo `ARIMA.predict` della classe `statsmodels.tsa.arima_model.ARIMA` che permette di fare la previsione in-sample.

Per quanto riguarda i test effettuati sono stati usati i seguenti metodi:

- Metodo `jarque_bera` della classe `statsmodels.stats.stattools` usato per effettuare il test di normalità di Jarque-Brera.
- Metodo `lilliefors` della classe `statsmodels.stats.diagnostic` usato per effettuare il test di normalità di Lilliefors.
- Metodo `adfuller` della classe `statsmodels.tsa.stattools` usato per effettuare il test di stazionarietà Augmented Dickey–Fuller.
- Metodo `test_heteroskedasticity` della classe `statsmodels.tsa.arima.model.ARIMAResults` usato per effettuare il test di eteroschedasticità (sottopopolazioni con diverse varianze).
- Metodo `acorr_ljungbox` della classe `statsmodels.stats.diagnostic` usato per effettuare il test di Ljung-Box dell'autocorrelazione nei residui.

Tutti questi test sono simili a quelli presenti in MATLAB.

Per quanto riguarda l'autocorrelazione dei residui è stata usata la funzione `acf` presente nella classe `from statsmodels.tsa.stattools`.

Capitolo 9

Conclusioni

9.1 Criticità

L'obiettivo iniziale dell'elaborato era quello di individuare un modello previsionale in grado di avere elevate performance in qualsiasi posizione geografica in cui fossero presenti informazioni di emissione. Tuttavia, la forte correlazione dovuta a determinismo dei dati ha comportato grosse difficoltà in fase di selezione del modello e, principalmente, nell'applicazione dei metodi statistici visti.
Il modello ottenuto per la serie storica nel punto di massima emissione che, come si è visto nel capitolo 7 è applicabile anche per i punti geografici classificati come emissioni medie, ha un'elevata capacità previsionale. La serie storica può essere descritta esclusivamente da modelli autoregressivi stagionali, senza l'intervento di regressori e senza l'ausilio di componenti a media mobile. La stagionalità è fortemente presente e si è visto come essa venga catturata ponendo il coefficiente autoregressivo a lag 12 pari a 1. Ciò evidenzia la forte dipendenza di ogni valore all'istante temporale esattamente un anno prima.
Allo stesso modo, si sono riscontrati risultati anomali nell'analisi della serie storica in termini di autocorrelazione tramite lag plot e correlogramma, con correlazioni forti e protratte nel lungo termine.
In fase di selezione del modello lo studio si è concentrato anche nell'esplorazione di modelli subottimi secondo la misura di performance principale adottata (AIC). Tuttavia, anche modelli semplici di ordini inferiori erano in grado di fittare il modello in un modo tale da avere residui fortemente correlati e concentrati attorno al valore zero, comportando una distribuzione fortemente non gaussiana.
Si è cercato quindi di ottenere risultati migliori sui residui, provando modelli di regressione con armoniche che, come atteso, han dato luogo a residui meno concentrati attorno al valore zero ma pur sempre eteroschedastici, non normali ed ancora più fortemente correlati.
Il modello ottenuto ha quindi una capacità inferiore di catturare informazioni dalla serie

storica presentata.

Infine, un ulteriore problema individuato in fase di selezione del modello è la particolare sensibilità del modello stesso alla porzione di dataset utilizzato in fase di training. Infatti, se le tecniche di cross validazione random e rolling aiutano nell'individuazione del modello che generalizzi quanto più possibile e sia meno dipendente dai dati utilizzati, nel processo di re-training finale si ottengono risultati significativamente diversi in funzione della dimensione e soprattutto dei primi dati selezionati della serie storica.
Nell'esempio mostrato in Figura 9.1, la suddivisione del dataset di training e validazione segue sempre una proporzione 70-30 ma vengono semplicemente esclusi i primi 12 mesi.

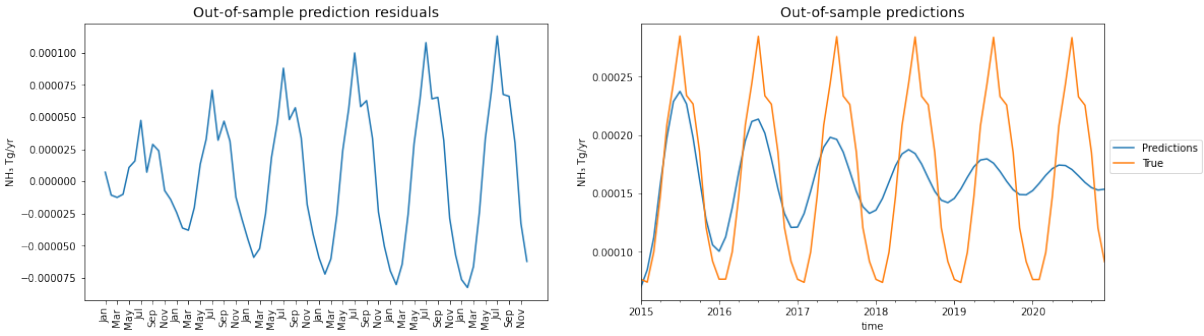

Figura 9.1: Previsioni errate a fronte di un diverso dataset di training

Il risultato è particolarmente strano, tanto che l'unica spiegazione ipotizzata sia la dipendenza dai dati iniziali dovuti al filtro di Kalman applicato in fase di fit del modello. Come si può notare nella Figura 2.7, la dipendenza tra il terzo anno (2002) ed il secondo anno (2001), nonché anno iniziale essendo il 2000 escluso, non è forte tanto quanto quelle degli anni successivi, con un conseguente coefficiente della componente autoregressiva inferiore ad 1, possibile causa del continuo smorzamento delle previsioni in figura.

9.2 Sviluppi futuri

Per prima cosa, si potrebbe procedere con un'analisi più dettagliata per l'individuazione del modello migliore nel punto di minimo, cercando di capire se le motivazioni che hanno portato ad un risultato come quello mostrato siano esclusivamente legati a problemi di scala molto piccola o ad altro. Un altro possibile approccio, con il solo fine ultimo di poter applicare metodi statistici a questa serie temporale deterministica, potrebbe essere quello di "sporcare" la serie storica con errori campionati da una distribuzione normale a media zero e con una data varianza. L'analisi risulterebbe fittizia in termini di andamento delle emissioni effettivamente stimate dall'organismo fornitore dei dati, ma potrebbe permettere di ottenere risultati più apprezzabili in termini statistici.