



UNIVERSITÀ
DEGLI STUDI
DI BERGAMO

Report progetto Statistica II

G12 Memphis

Cambiamenti delle fonti di energia tra il XX e XXI secolo
negli USA e analisi delle emissioni di CO₂

Matteo Ballabio - 1058828

Matteo De Stefani - 1080379

Lisa Galizzi - 1059947

Vanessa Zani - 1057577

1 Introduzione

1.1 Background e razionale

La globalizzazione e la crescita del capitalismo hanno portato ad un inevitabile aumento della produzione e ciò ha avuto un impatto negativo sull'ambiente. A fronte di ciò uno dei temi più dibattuti negli ultimi anni è stato, e continua ad essere, il riscaldamento globale: la CO₂ e gli altri gas ad effetto serra si accumulano nell'atmosfera alterandone la composizione chimica, in conseguenza della loro capacità di trattenere calore il loro accumulo provoca un aumento dell'effetto serra naturale, generando un riscaldamento del clima terrestre. Tutti i paesi del mondo si stanno attivando per cercare soluzioni per impedire la realizzazione dello scenario futuro che attualmente si prospetta, caratterizzato da un cambiamento radicale degli ecosistemi con relativa estinzione delle specie che li popolano.

Il presente lavoro è incentrato sul territorio degli Stati Uniti che costituisce uno tra i paesi che attualmente producono più emissioni di CO₂ e uno dei leader più potenti a livello mondiale dal punto di vista economico e di consumo di energia. È opportuno in questo contesto citare le COP (Conferences of the Parties), conferenze sul clima alla quale prendono parte i paesi che hanno ratificato la Convenzione Quadro delle Nazioni Unite sui Cambiamenti Climatici, tra cui gli USA.

Il primo documento internazionale che ha imposto dei limiti sulle emissioni di CO₂ ai paesi più ricchi viene stipulato nel 1997 durante la COP3: il Protocollo di Kyoto. Il documento stabilì che per gli USA le emissioni dovevano essere ridotte del -7%, rispetto ai livelli del 1990, dal 2008 al 2012. Nel 2001 tuttavia gli USA si ritirano dal Protocollo per volere dell'allora presidente George W. Bush. Altra tappa importante nella storia delle COP si ebbe durante la COP21, nel 2015, nella quale si raggiunse l'Accordo di Parigi che stabilì di limitare il riscaldamento globale a +1,5° C; l'accordo è entrato in vigore nel 2016 durante la presidenza di Barack Obama, il quale accolse positivamente la proposta. Solo un anno più tardi il nuovo presidente Trump annunciò il ritiro degli USA anche da questo accordo considerando gli obiettivi in esso stabiliti "non realistici e con maggiori costi che benefici"; il paese oltreoceano sotto la presidenza di Trump rimase quindi totalmente indifferente al problema delle emissioni. Il 26esimo incontro si è tenuto quest'anno e sia gli USA che la Cina, altro colosso mondiale a livello di emissioni, hanno dichiarato l'intenzione di lavorare insieme per raggiungere l'obiettivo di contenere l'aumento della temperatura media globale non oltre i +1,5 °C rispetto al periodo pre-industriale, come stabilito dagli accordi scaturiti dalla COP21. Si prospettano cambiamenti, per quanto possibile, positivi poiché l'attuale presidente Joe Biden si dichiara sensibile al tema del cambiamento climatico.

Attualmente le fonti principali delle emissioni sono gli impianti che producono energia e la circolazione di mezzi pesanti e al momento l'arma migliore per contrastarle è fare sempre più affidamento sulle fonti rinnovabili. I loro costi sono scesi e ci sono metodi sempre nuovi per integrarle; ciò significa che maggiori quantità di energia verranno prodotte direttamente negli Stati Uniti invece di essere importate e che le emissioni di CO₂ caleranno. I benefici derivanti dalla diminuzione delle emissioni di CO₂ sono molteplici: una migliore qualità dell'aria e riduzione di morti, malattie e perdite economiche legate all'inquinamento. (Shindell et al. (2021))

Produzione e consumo di energia non rinnovabile

Nella figura a seguito, l'andamento colorato in blu corrisponde in parte alla somma dei tre andamenti nella parte bassa del grafico e rappresenta la produzione totale di fonti di energia non rinnovabili. Allo stesso modo l'andamento in verde, come indicato in legenda, mostra l'andamento del consumo delle fonti di energia non rinnovabili.

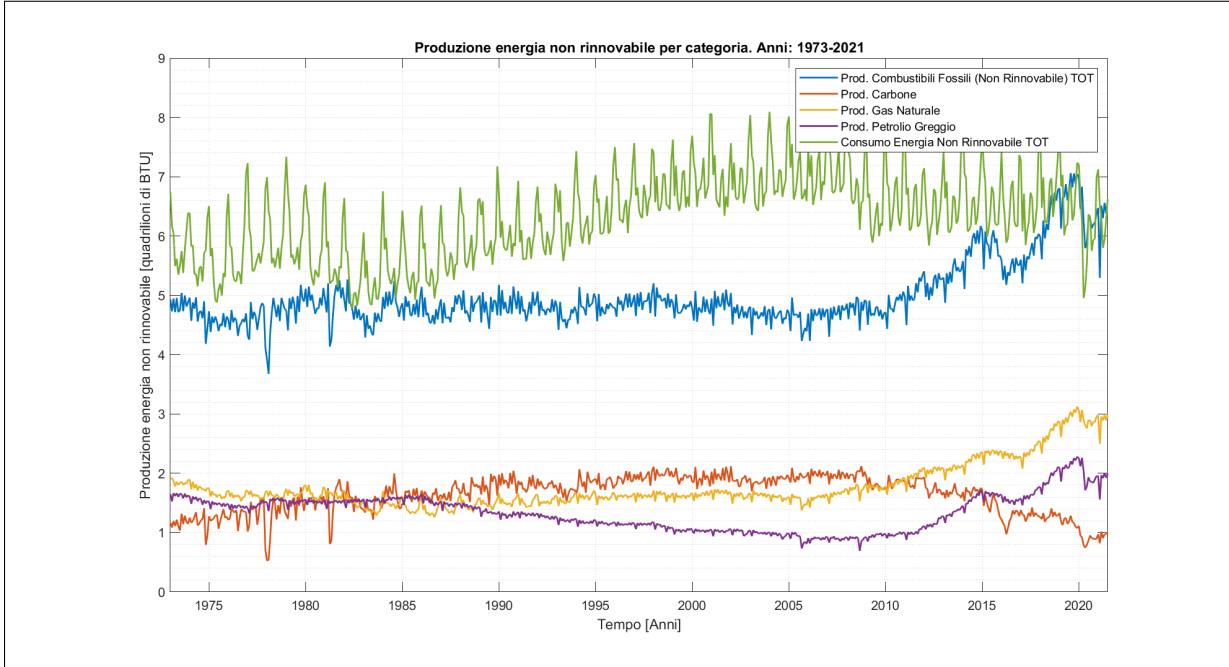


Figura 1: Serie storiche energie non rinnovabili

È immediato notare che il consumo è prevalentemente superiore alla produzione e per sopperire a questa mancanza gli USA ricorrono all'importazione di combustibili fossili da paesi esteri membri dell'Organizzazione dei Paesi Esportatori di petrolio (OPEC). Negli ultimi anni invece la produzione supera il consumo e ciò è in parte dovuto all'introduzione di fonti di energia rinnovabili. Tra le fonti di energia non rinnovabili abbiamo scelto di focalizzarci su quelle che ci sembravano più significative: petrolio greggio, carbone e gas naturale. Da una prima analisi è possibile notare che dal 1975 al 2010 sono visibili andamenti più o meno regolari, tranne per quanto riguarda quello della produzione del petrolio greggio che ha subito una lenta decrescita. Dal 2010 ai giorni nostri è in calo la produzione di carbone in quanto il suo utilizzo, soprattutto per produrre elettricità, è stato sostituito dall'utilizzo di liquidi derivanti dal petrolio e dall'utilizzo di fonti rinnovabili. La produzione di petrolio greggio dal 2010 ad oggi è quasi duplicata; gli Stati Uniti sono infatti il terzo produttore al mondo di greggio e sono allo stesso tempo il paese che consuma più petrolio al mondo: esso infatti copre circa il 35% della domanda interna di energia e viene utilizzato soprattutto nel settore dei trasporti e dell'industria. Anche per quanto riguarda la produzione di gas naturale è possibile notare che dal 2010 ad oggi l'andamento è in crescita e ciò è dovuto all'uso di tecniche di trivellazione e di produzione più efficienti (che tuttavia producono inquinamento atmosferico). Il gas naturale produce meno polveri sottili del carbone e dei prodotti petroliferi raffinati e ciò ha contribuito a un suo maggiore utilizzo nella produzione di elettricità e nel settore dei trasporti. Tuttavia, il gas naturale è fatto per lo più di metano, un potente gas a effetto serra che si disperde nell'atmosfera.

Produzione e consumo di energia rinnovabile

Gli USA si sono portati sul podio, insieme alla Cina e all’Australia, per essere i paesi con maggior produzione di energia rinnovabile; dal grafico a seguito è immediato notare che si tratta di quantità di energia minori rispetto a quelle riscontrabili nel grafico precedente relativo alle produzioni e al consumo di fonti di energia non rinnovabili.

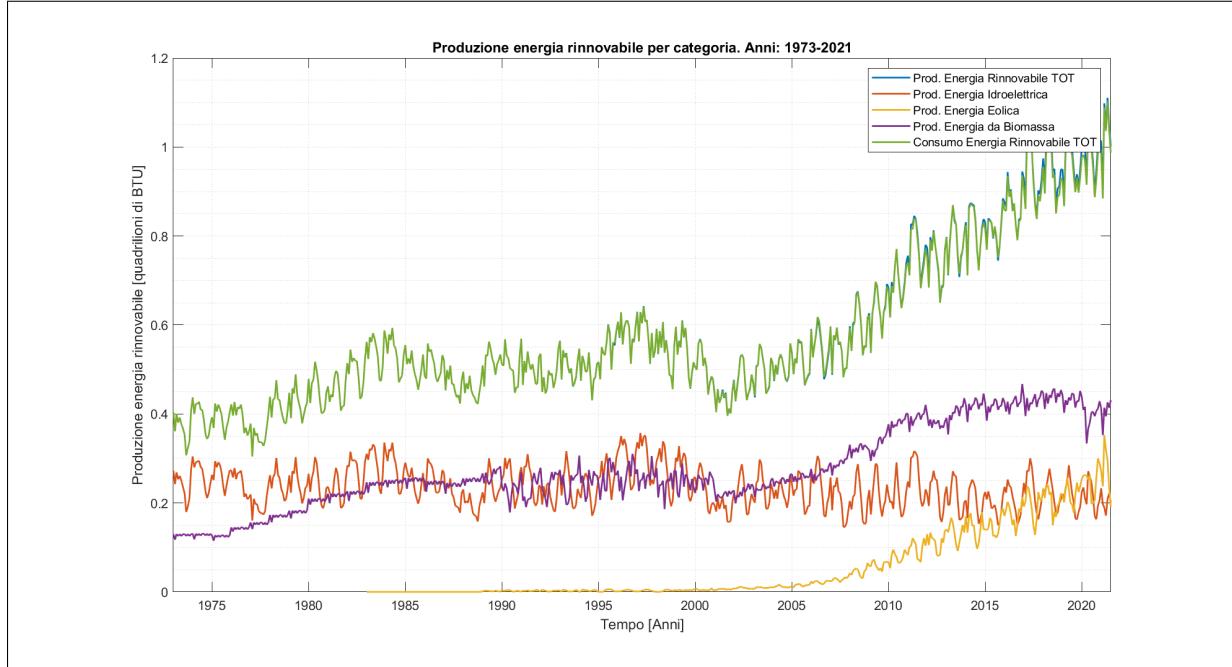


Figura 2: Serie storiche energie rinnovabili

In questo caso gli andamenti della produzione totale e del consumo totale di energia rinnovabile sono quasi totalmente sovrapposti e ciò è dovuto al fatto che i quantitativi prodotti sono relativamente bassi e vengono utilizzati quasi interamente contestualmente alla produzione. I tre andamenti visibili nella parte bassa del grafico corrispondono agli andamenti delle tre fonti di energia più significative: idroelettrica, eolica e da biomasse. L’energia idroelettrica corrisponde ad una delle prime fonti di energia rinnovabile introdotta a livello globale e l’andamento negli anni è rimasto più o meno costante. L’energia eolica è una fonte che è stata introdotta in modo sostanzioso solo recentemente, ha subito un’esplosione a partire dal 2007 con un picco massimo registrato proprio nell’ultimo anno; sembra quasi un paradosso che tale risultato sia stato raggiunto proprio al termine della discussa presidenza Trump, che di certo non ha fatto dell’ambientalismo un punto cardine vincolante nelle scelte economico-politiche. Per ultimo nel grafico è mostrato l’andamento della produzione di energia da biomasse che viene principalmente usata per produrre calore e vapore per l’industria o per il riscaldamento; tra esse si annoverano biocombustibili come etanolo e biodiesel, che si propongono di sostituire l’utilizzo di combustibili fossili nel settore dei trasporti in quanto presentano il pregio di ridurre le emissioni di CO₂ in atmosfera. Essi sono fortemente promossi da incentivi statali e del governo federale per questo si stima che nei prossimi anni il loro uso sarà in crescita.

Emissioni di C0₂ per differenti combustibili fossili

Gli USA attualmente emettono il 15% delle emissioni di carbonio di tutto il mondo e il presente grafico mostra come i livelli di emissioni di C0₂ totali sono cambiati nel corso degli anni. Il trend è attualmente in discesa e la motivazione principale è l'incremento nell'utilizzo di fonti di energia rinnovabili.

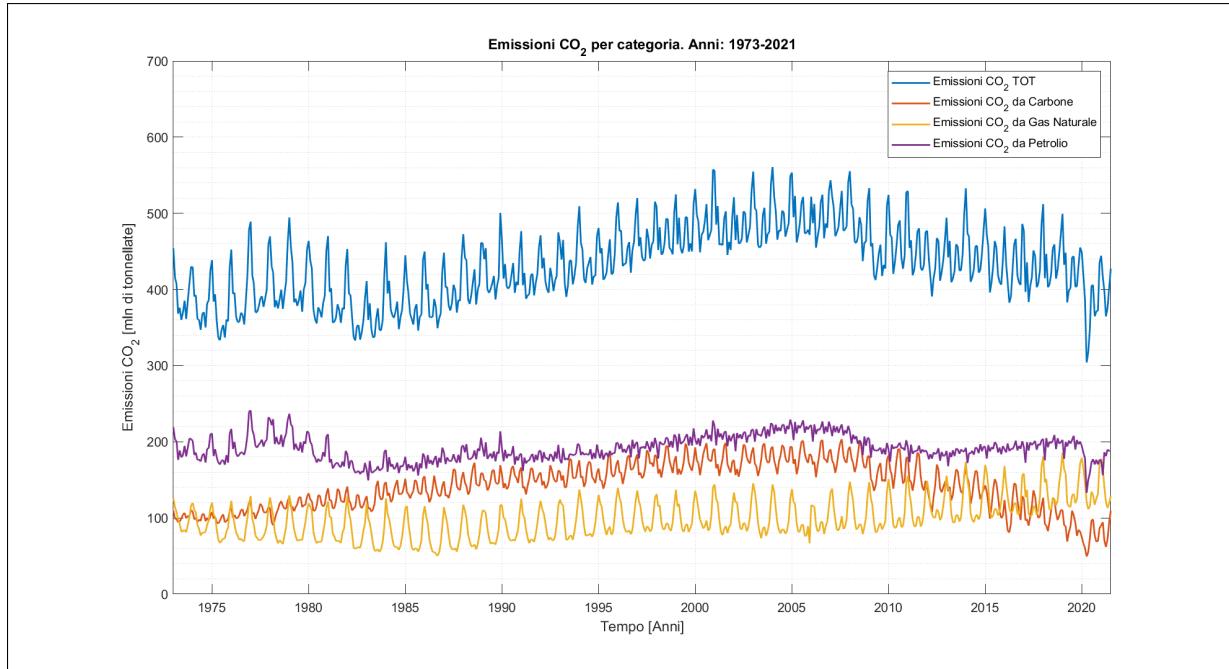


Figura 3: Serie storiche confronto emissioni di C0₂

1.2 Obiettivi del progetto

Il presente studio si focalizza sulla stima delle emissioni di C0₂ negli USA; partendo da un'analisi di correlazione semplice fra singole variabili è stato possibile costruire un modello in grado di mimare l'andamento delle emissioni grazie ai dati relativi ai consumi e alle produzioni delle varie fonti di energia elencate nel dataset. Successivamente si è passato a modelli previsivi sempre basati sulle variabili relative a produzioni e consumi e sono stati confrontati diversi modelli per vedere quale fosse il migliore. In seguito, vista la relazione tra il clima e le emissioni di CO₂ si è cercato di costruire un modello relativo a quest'ultime servendoci di variabili climatiche quali HDD e anomalie sulle temperature. Infine è stato possibile verificare come le anomalie sulle temperature globali siano influenzate dalle emissioni di C0₂ dei tre colossi USA, Cina e Russia.

2 Stima delle emissioni di CO₂ basata su produzioni e consumi di fonti energetiche

Analisi della distribuzione

Per stimare le emissioni di CO₂ è stata utilizzata una parte del dataset mensile, in particolare si è preso come riferimento la serie storica a partire da Gennaio 2010, cioè si esegue una stima con dati mensili comprendenti un arco temporale di 11 anni.

Prima di effettuare la stima è bene plottare i dati per vedere se l'andamento è approssimabile ad una normale, cioè che abbia la seguente relazione:

$$f_{\mu, \sigma^2}(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right)$$

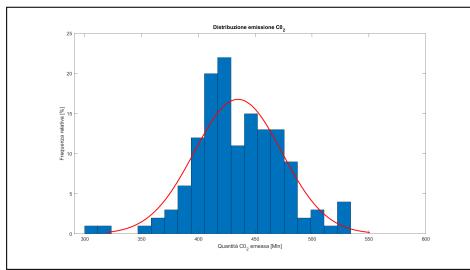


Figura 4: Curva di distribuzione delle emissioni di CO₂

In particolare, si possono ora analizzare 2 parametri in riferimento alla curva rappresentata nella figura 4:

- Simmetria: considerando la simmetria della curva, cioè quanto è centrata rispetto alla media μ , si utilizza l'indice skewness:

$$sk = E\left[\left(\frac{y-\mu}{\sigma}\right)^3\right]$$

Il risultato ottenuto è stato: $sk = -0.1005 < 0$. Questo valore, valutato rispetto allo 0, indica che la curva non è perfettamente centrata in $\mu = 0$, ma si trova alla sua sinistra, quindi si ha una lieve asimmetria negativa;

- Ampiezza: si utilizza l'indice curtosi, che rappresenta la nitidezza relativa del picco della curva di distribuzione di probabilità, in altre parole accerta il modo in cui le osservazioni sono raggruppate attorno al centro della distribuzione:

$$k = E\left[\left(\frac{y-\mu}{\sigma}\right)^4\right]$$

Il risultato ottenuto è stato: $k = 3.8799 > 3$. Questo valore deve essere confrontato rispetto al valore di riferimento 3: in questo caso l'indice di curtosi è maggiore, quindi ciò indica che vi è un picco centrale rispetto alla distribuzione normale.

Entrambi gli indici dimostrano quanto è visibile dal grafico 4: la curva risulta infatti essere decentrata rispetto alla media, con un picco maggiore spostato verso sinistra.

Test di normalità

- **BERA – JARQUE:** si consideri l'ipotesi nulla (H_0) che la distribuzione della popolazione sia normale, contro l'ipotesi H_1 di non normalità dei dati. Il Test di Normalità di Jarque-Bera è basato sulla vicinanza dell'asimmetria campionaria a 0 e della curtosi campionaria a 3.

$$JB_n = (n/6)(sk)^2 + n/24(\hat{k} - 3)^2$$

Per determinare se accettare o rifiutare H_0 si deve confrontare il p-value ottenuto con ogni livello di significatività considerata, accettando H_0 quando $p > \alpha$, rifiutandola altrimenti.

In questo caso il test ha prodotto un p-value: $p = 0.0689$ Utilizzando 3 diversi valori di significatività del test (α), rispettivamente 0.1, 0.05 e 0.01, la distribuzione risulta normale per $\alpha = 0.1$ e $\alpha = 0.05$.

- **LILLIEFOR:** come nel caso del test di Bera – Jarque, se $p < \alpha$ si rifiuta l'ipotesi nulla.

$$D = \max_{t=1,\dots,n} \left| \hat{F}(y_t) - \phi\left(\frac{y_t - \bar{y}}{s_y}\right) \right|$$

Il test ha prodotto un p-value: $p = 0.1486$. Con un livello di significatività del 5% si accetta l'ipotesi di normalità.

Correlazione

La correlazione, misurata attraverso il coefficiente di correlazione (r), indica come una o più variabili cambiano rispetto alla variabile indipendente considerata. In base al valore del coefficiente r ($-1 \leq r \leq 1$) è possibile capire se tra le variabili vi sia una relazione lineare e quanto quest'ultima sia forte: quanto più $|r| = 1$, tanto maggiore sarà la relazione lineare (positiva o negativa) tra i parametri presi in riferimento.

La relazione di correlazione è espressa dalla seguente formula:

$$Y = X\beta + \epsilon$$

Per decidere quali variabili correlare si utilizza uno scatter plot in cui vengono indicati i coefficienti di correlazione:

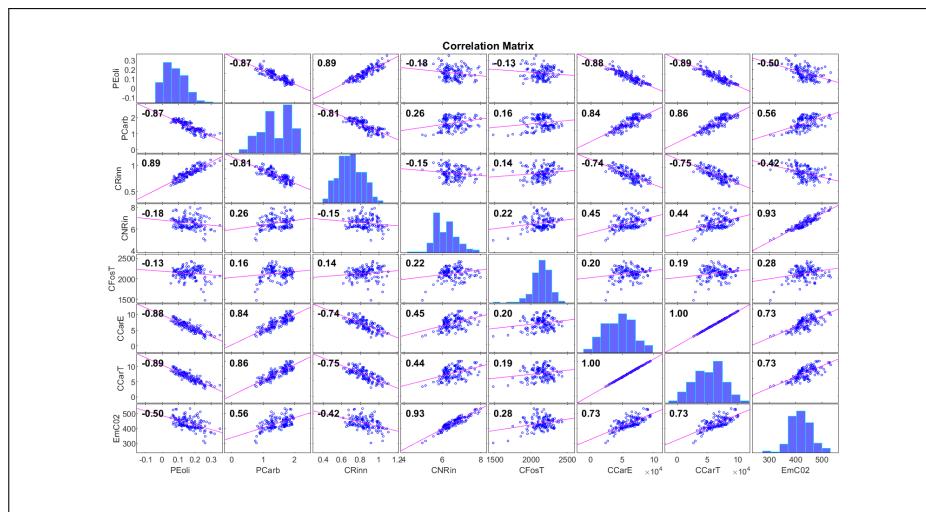


Figura 5: Scatter plot al fine di valutare le correlazioni con le emissioni di CO₂

Regressione lineare semplice

Modello

Considerando la matrice in figura 5, è evidente che la retta che rappresenta una buona correlazione lineare tra emissione di CO₂ e produzione, si ha con la produzione totale del carbone. In particolare vi è una correlazione dello 0.73.

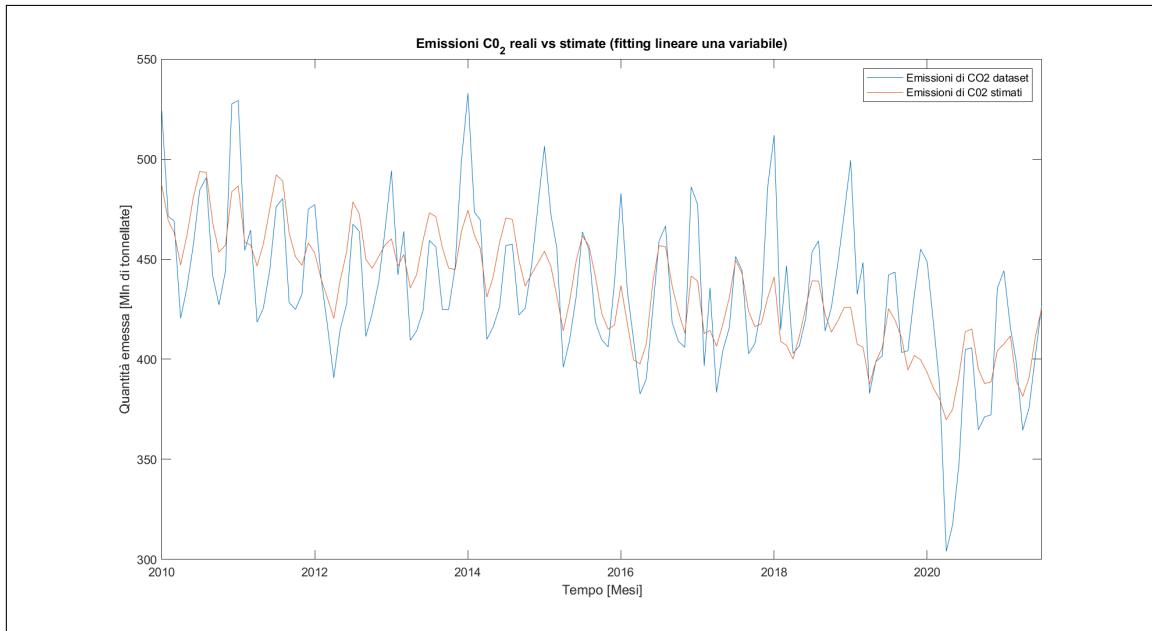


Figura 6: Serie storica delle emissioni di CO₂ con stima basata su una variabile

I risultati ottenuti mostrano un'intercetta significativa ad ogni livello di significatività ($p < 0.01$), così come il consumo totale di carbone risulta significativo ad ogni livello di significatività.

L'errore al quadrato risulta essere il seguente: $R^2 = 0.537$, ciò indica che il modello trovato non è molto significativo utilizzando solo la variabile della produzione totale del carbone.

Dal grafico 6 si nota che la stima fatta attraverso una regressione semplice non è in grado di stimare il corretto andamento dell'emissione di CO₂, è in grado di prevedere i picchi e i minimi dell'andamento, ma non di stimarli nel modo corretto in quanto, a parte per Luglio 2015 e 2017 in cui i massimi delle due curve si sovrappongono, nella maggior parte dei casi il modello sottostima i massimi e i minimi e non è stato in grado di prevedere il crollo di emissioni durante il lockdown.

Analisi dei residui

I residui rappresentano le differenze tra i valori osservati nel dataset e i valori stimati calcolati con l'equazione di regressione; essi indicano la variabilità dei dati attorno alla retta di regressione e quindi la parte di errore di previsione del modello.

Il termine d'errore (ϵ) che appare nella formula della regressione semplice deve essere una variazione imprevedibile nella variabile risposta, in modo che il modello di regressione riesca ad avere un buon potere predittivo. Per verificare se è effettivamente così, quando si costruisce un modello di regressione bisogna verificare la distribuzione dei residui.

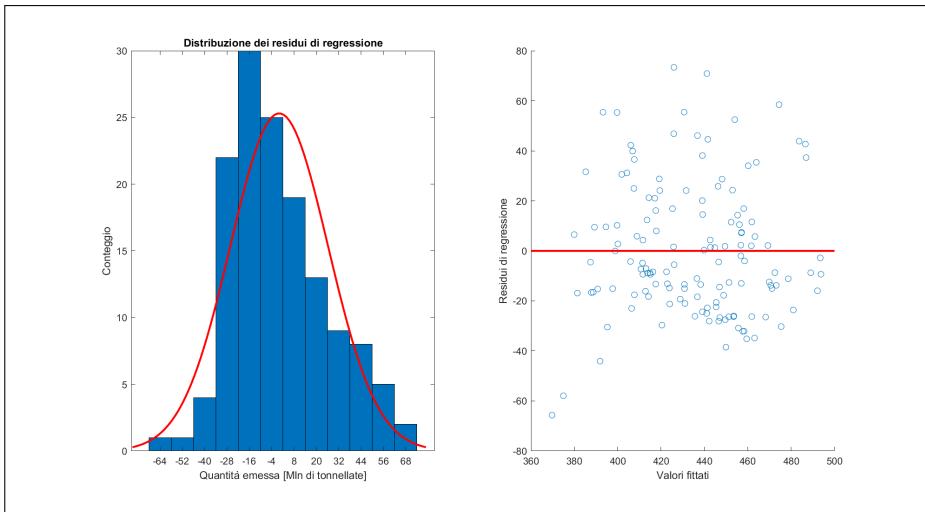


Figura 7: Analisi della distribuzione residui e scatter plot

Attraverso gli indici di normalità skewness e curtosi e i test di normalità svolti sui residui, si può concludere che la distribuzione è riconducibile ad una gaussiana.

Poiché la stima eseguita con la regressione semplice ha prodotto un risultato troppo approssimativo, si è deciso di applicare una regressione utilizzando più variabili, in modo da avvicinarci alla stima reale.

Regressione lineare multipla

Modello con predittori totali

Variabili utilizzate: produzione energia rinnovabile totale, consumo energia non rinnovabile totale e consumo energia rinnovabile totale.

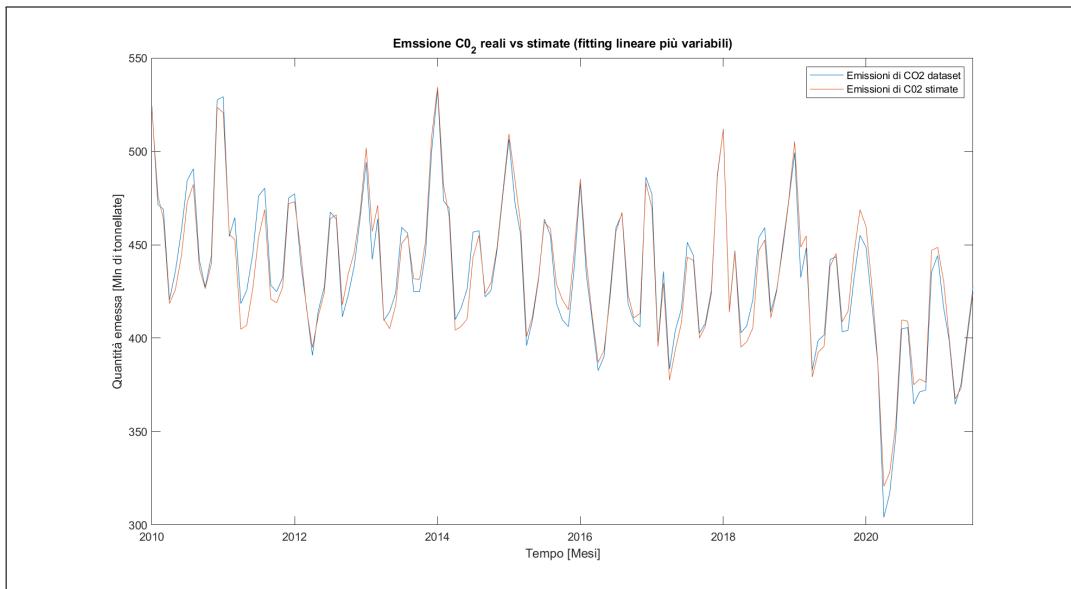


Figura 8: Serie storica delle emissioni di CO₂ con stima basata su più variabili

La figura 8 mostra che il modello con la regressione lineare multipla è in grado di stimare l'andamento reale con un alto grado di attendibilità, infatti vi è un R^2 di 0.961. Il modello non solo si sovrappone per buona parte dell'andamento al grafico delle emissioni reali, ma è stato in grado di prevederne anche il crollo registrato durante il primo lockdown di Aprile 2020.

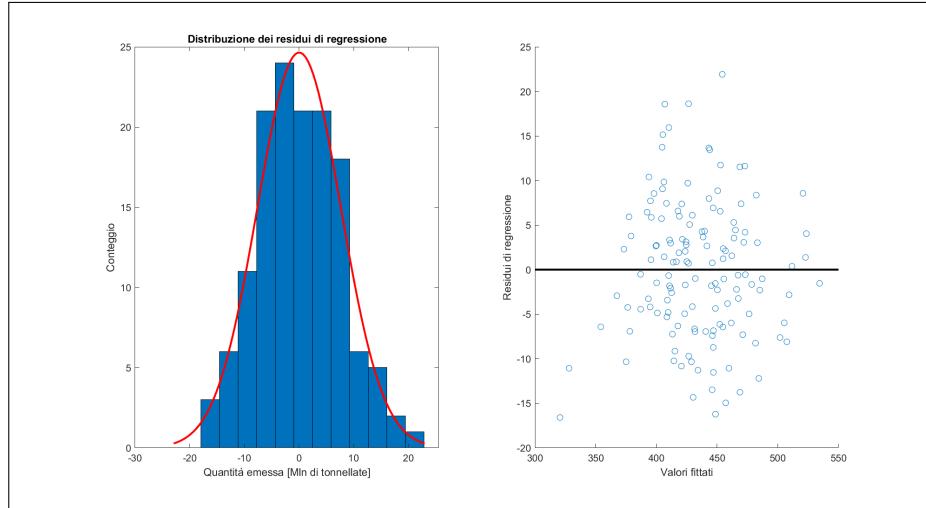


Figura 9: Analisi della distribuzione residui e scatter plot

Attraverso gli indici di normalità skewness e curtosi e i test di normalità svolti sui residui, si può concludere che la distribuzione è riconducibile ad una gaussiana e che il modello è migliore del precedente.

Modello con predittori singoli

Variabili utilizzate: produzione eolica, produzione carbone, consumi del carbon fossile nel settore dei trasporti, produzione biomasse, consumo carbone totale.

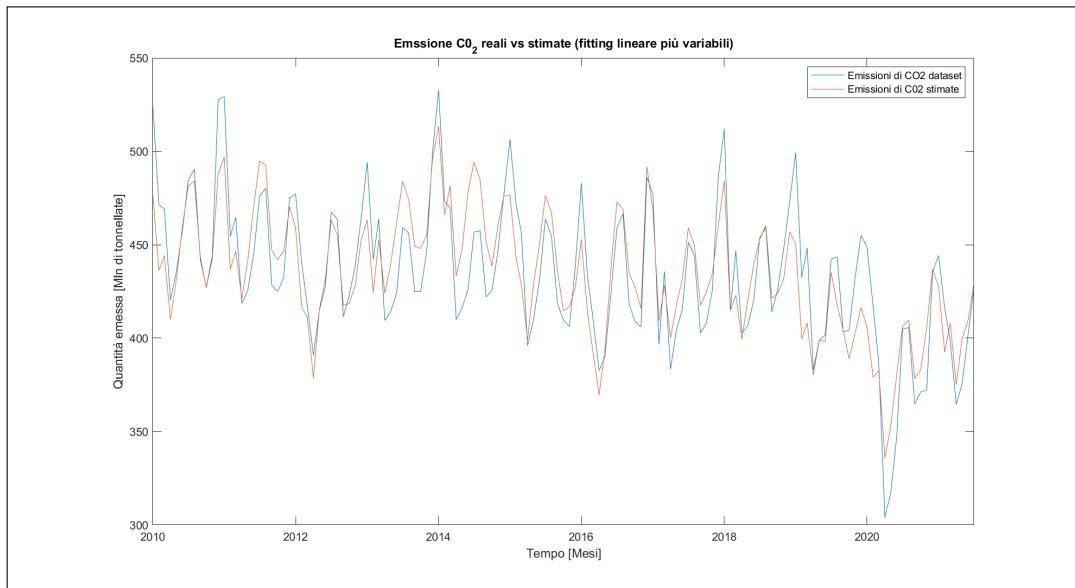


Figura 10: Serie storica delle emissioni di CO₂ con stima basata su più variabili

Anziché utilizzare come predittori i totali dei consumi e delle produzioni, sono stati presi in riferimento i dati singoli, senza considerare il totale.

Il risultato è il seguente: $R^2 = 0.731$: migliora il modello della regressione lineare semplice, ma non quello in cui si sono utilizzati i dati aggregati. Per aumentare R^2 occorre fare una selezione delle variabili utilizzando metodi più accurati, che permettono di stimare un modello che rappresenta meglio la realtà.

Regressione con metodo Stepwise

La stepwise è un metodo di selezione delle variabili indipendenti allo scopo di selezionare un set di predittori che abbiano la migliore relazione con la variabile dipendente. Esistono 2 approcci di selezione delle variabili:

- Il metodo forward (in avanti) inizia con un modello vuoto nel quale nessuna variabile tra i predittori è selezionata; nel primo step viene aggiunta la variabile con l'associazione maggiormente significativa sul piano statistico. Ad ogni step successivo è aggiunta la variabile con la maggiore associazione statisticamente significativa tra quelle non ancora incluse nel modello e il processo prosegue sino a quando non vi è più variabile con associazione statisticamente significativa con la variabile dipendente;
- Il metodo backward (all'indietro) inizia con un modello che comprende tutte le variabili e procede, step by step, ad eliminare le variabili partendo da quella con l'associazione con la variabile dipendente meno significativa sul piano statistico.

Il processo stepwise fa avanti e indietro tra i due processi, aggiungendo e rimuovendo le variabili che, nei vari aggiustamenti del modello (con aggiunta o re-inserimento di una variabile) guadagnano o perdono in termini di significatività.

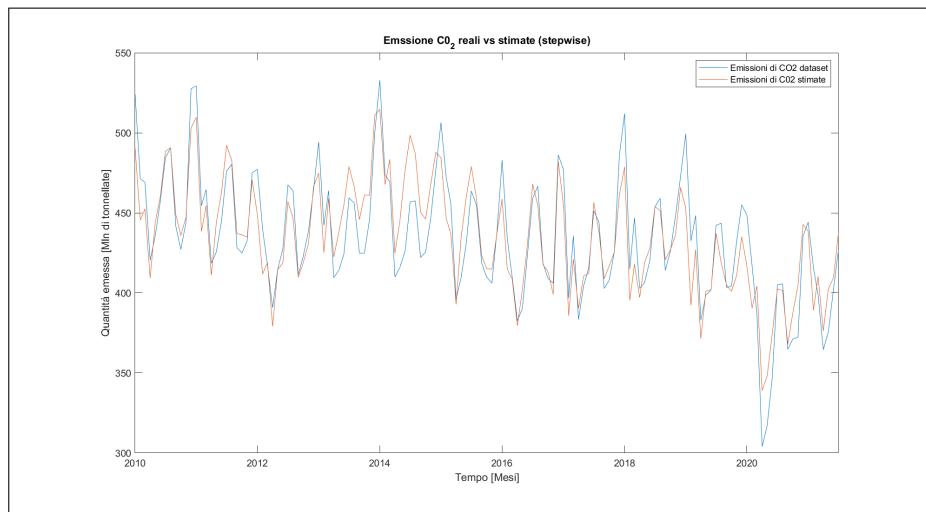


Figura 11: Serie storica delle emissioni di CO₂ con stima basata su stepwise

Regressione con metodo Lasso

Il Lasso è un algoritmo più complesso e solitamente più preciso rispetto alla stepwise, in quanto valuta ogni variabile, conferendo a ciascuna di essa un peso diverso in base al coefficiente di regressione stimato. L'algoritmo si compone essenzialmente in 2 parti:

- Inizialmente valuta i coefficienti di regressione e scarta quelli che vanno a 0 in quanto sono considerati non significativi (figura 13);
- Successivamente pesa i coefficienti rimasti in base al livello di correlazione: se sono fortemente correlati (positivamente o negativamente) alla variabile risposta avranno un peso maggiore e saranno quindi tenuti in maggior considerazione perché sono quelli più significativi per la stima del modello.

$$\hat{\beta}_\lambda = \arg \min_{\beta} \sum_{t=1}^n (y_t - \hat{y}_t)^2 + \lambda \sum_{j=1}^k |\beta_j|$$

La formula del Lasso è caratterizzata da una parte iniziale, in cui si minimizza la somma quadrata degli errori, mentre la seconda considera diversi valori del parametro λ , che è il fattore di controllo della contrazione, e seleziona il valore minore in cross-validation (figura 12).

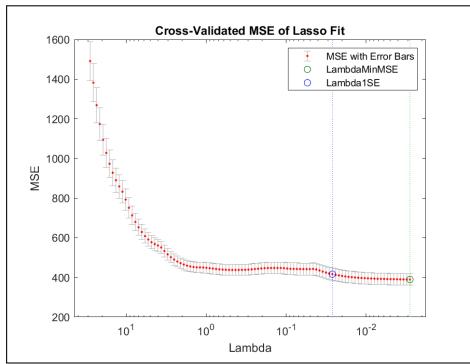


Figura 12: Andamento λ in cross-validazione

Quando λ è uguale a zero, il modello diventa la regressione dei minimi quadrati ordinari. Di conseguenza, quando λ aumenta, la varianza diminuisce in modo significativo e aumenta anche la distorsione nel risultato, quindi i coefficienti sono forzati a essere zero perché non sono significativi.

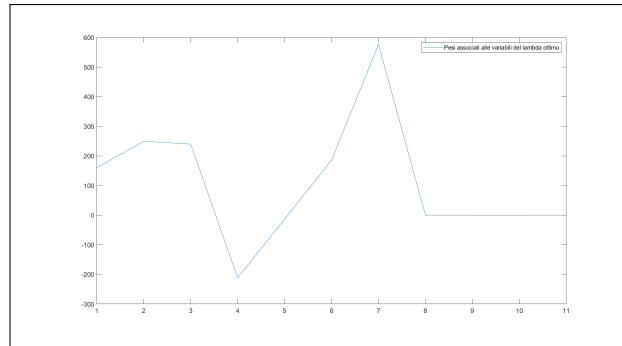


Figura 13: Incidenza dei coefficienti di regressione

Le variabili sono state ottimizzate mediante il metodo Lasso e sono stati assegnati i rispettivi pesi. In particolare, le ultime 3 variabili sono state portate ad un valore molto vicino a 0 (figura 13).

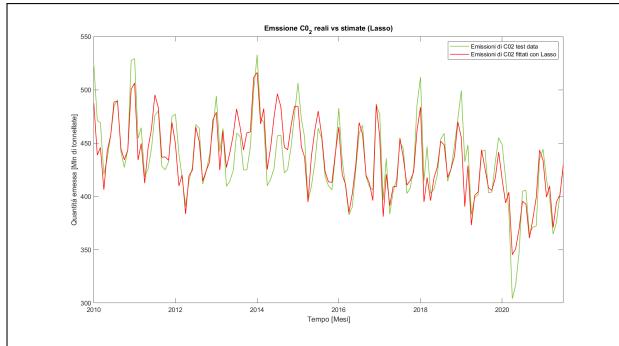


Figura 14: Serie storica delle emissioni CO₂ vs stimate con Lasso

Il modello generato attraverso il Lasso ha prodotto un R^2 di 0.794, quindi vi è un miglioramento rispetto alla curva delineata grazie alla funzione stepwise, che dava in output un R^2 di 0.785.

	R^2	sk	k	pv B-J ($\alpha = 5\%$)	pv Lilliefors ($\alpha = 5\%$)
Reg. semplice	0.537	0.5687	3.0127	0.0284	1.0000e – 03
Reg. mult. (aggr.)	0.961	0.2170	2.8723	0.5	0.5
Reg. mult. (sing.)	0.731	0.2624	2.6842	0.2678	0.2249
Stepwise (sing.)	0.785	–	–	–	–
Lasso (sing.)	0.794		–	–	–

In conclusione, osservando la tabella soprastante, si può affermare che il modello che produce una stima migliore dell'emissione di CO₂ è quello basato sulla regressione multipla con variabili aggregate, in quanto ha un R^2 maggiore. Considerando i modelli stimati con la regressione su variabili singole, la curva migliore è quella descritta con il Lasso ($R^2 = 0.794$).

3 Previsione delle emissioni di CO₂ basata su modelli ARIMA

Introduzione

I modelli della famiglia ARIMA permettono di ottenere una previsione a breve termine di serie storiche. L'utilizzo di questi modelli permette di modellare una serie storica basandosi sull'autocorrelazione totale e parziale della serie indagata. Nel nostro caso, si è applicato un modello di questo tipo per la previsione delle emissioni di CO₂ negli USA da Gennaio 2010 a Luglio del 2021. E' stato scelto questo range temporale in quanto nella serie sono presenti due fenomeni interessanti: la diminuzione graduale ma costante delle emissioni nell'arco temporale e il fenomeno esogeno del COVID-19 (Marzo-Aprile 2020).

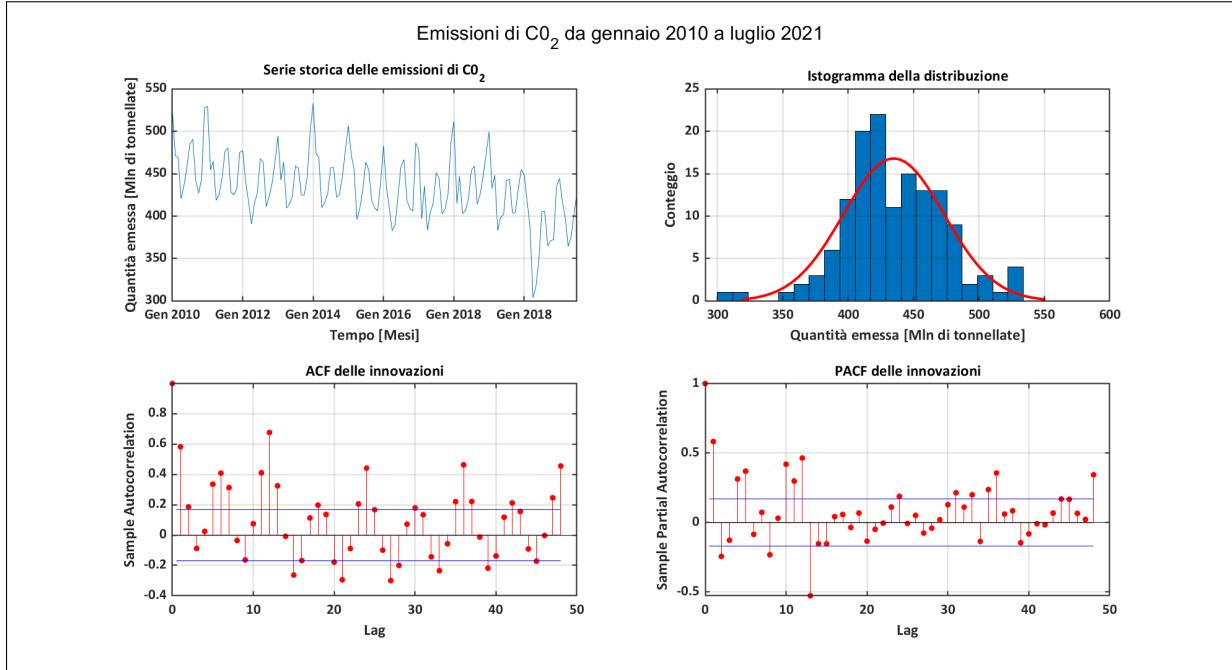


Figura 15: Tabella riassuntiva con: a) serie storica b) istogramma serie storica c) ACF d) PACF

Il primo passaggio è stata la valutazione della serie storica mediante l'utilizzo di un correlogramma delle autocorrelazioni totali e parziali. L'autocorrelazione e l'autocorrelazione parziale permettono di verificare la somiglianza interna della serie storica, ossia valutare se la serie traslata di un certo ritardo ($t + 1$) assomiglia all'istante precedente (t) (Dagum (2001)). Dalla valutazione qualitativa della serie storica, notiamo mesi con una certa autocorrelazione, in particolare Gennaio (ogni 12 mesi), Aprile e Ottobre (ogni 6 mesi). Nella maggior parte dei casi, si è valutato anche il PACF e l'autocorrelazione parziale più significativa (nel grafico linea rossa esterna alle bande blu) rimaneva a ritardo 12. Per questo motivo, la prima interpretazione è stata che almeno per i primi 48 mesi abbiamo una serie stagionale a ritardo 12. Questa valutazione è stata confermata mediante l'utilizzo di un test statistico che permette di quantificare l'autocorrelazione a differenti ritardi (Ljung-Box test). Le ipotesi del test sono le seguenti:

$$H_0 : \text{Non è presente autocorrelazione} \rightarrow \text{ipotesi nulla}$$

$$H_1 : \text{Presente autocorrelazione} \rightarrow \text{ipotesi alternativa}$$

Inoltre, si nota una certa stazionarietà della serie storica con un trend decrescente nell'arco temporale. Questa

ipotesi è stata verificata mediante il test statistico di Augmented-Dickey-Fuller.

H_0 : la serie è non stazionaria \rightarrow ipotesi nulla

H_1 : la serie è stazionaria \rightarrow ipotesi alternativa

L'ipotesi della presenza del trend lineare decrescente è stata confermata dal test statistico: per i primi 12 ritardi la serie è stazionaria con trend decrescente ($p\text{-value} < 5\%$). L'obiettivo dei modelli che saranno presentati nella sezione successiva è la previsione degli ultimi 24 mesi (da Luglio 2019 a Luglio 2021). In particolar modo, si è deciso di utilizzare l'85% del dataset ridotto di 11 anni per l'addestramento del modello (learning set) e il 15% per la valutazione del modello (test set). I valori fittati subiscono alcune perdite di performance dovute al fatto che la serie non è completamente stazionaria e poichè all'interno del test set è presente un fattore esogeno che influenza notevolmente la serie storica (COVID-19).

Modello AR(12)

Il primo modello costruito è un modello un po' grezzo che permette però di valutare quali ritardi sono significativi mediante l'utilizzo della sola parte autoregressiva. In particolar modo, l'equazione analitica che descrive il modello si può descrivere nel seguente modo:

$$y_t = a_1 y_{t-1} + a_2 y_{t-2} + \dots + a_{12} y_{t-12} + \varepsilon_t$$

I risultati del modello sono stati riassunti nella seguente tabella:

ARIMA(12,0,0) Model (Gaussian Distribution)				
	Value	StandardError	TStatistic	PValue
Constant	35.248	55.128	0.6394	0.52257
AR{1}	0.23349	0.07708	3.0292	0.0024523
AR{2}	0.0033727	0.08631	0.039076	0.96883
AR{3}	-0.040055	0.094621	-0.42332	0.67206
AR{4}	0.052005	0.11952	0.4351	0.66349
AR{5}	0.045041	0.12757	0.35306	0.72404
AR{6}	-0.08504	0.10688	-0.79569	0.42621
AR{7}	0.11521	0.10888	1.0581	0.29001
AR{8}	-0.088496	0.10584	-0.83615	0.40307
AR{9}	-0.13347	0.089312	-1.4944	0.13507
AR{10}	0.11149	0.078287	1.4241	0.15441
AR{11}	0.0022594	0.081469	0.027734	0.97787
AR{12}	0.70115	0.071708	9.7778	1.4017e-22
Variance	259.2	33.113	7.828	4.9576e-15

Figura 16: Tabella riassuntiva del modello AR(12)

Innanzitutto, si nota che gli unici due AR realmente significativi sono quello a ritardo 1 e 12 ($p\text{-value} < 0.001$). Questo suggerisce che è possibile valutare di ridurre la complessità del modello andando ad eliminare tutti gli AR intermedi e introducendo una stagionalità deterministica a ritardo 12. Due metodi di valutazione dei

modelli da valutare sono AIC e BIC. Il BIC rispetto all'AIC porta alla penalizzazione del modello per l'aumento dei parametri utilizzati. In particolar modo l'AIC risulta molto più basso, quindi più performante, rispetto al BIC proprio per l'aggiunta di molti parametri (ossia tutti gli AR intermedi).

Modello SAR(12)

Il modello seguente prevede l'assunzione di stagionalità della serie storica indagata, verificabile mediante il grafico delle autocorrelazione totali e parziali. In particolar modo, come detto prima, un'autocorrelazione a ritardo 12 suggerisce ad esempio che Gennaio del 2020 è fortemente autocorrelato con Gennaio 2019. Per questo motivo, mediante la minimizzazione di AIC e BIC, tenendo conto della componente stagionale a ritardo 12, il modello scelto è stato SAR(1,0,0)(12,0,0).

ARIMA(1,0,0) Model with Seasonal AR(12) (Gaussian Distribution)				
	Value	StandardError	TStatistic	PValue
Constant	23.473	9.2856	2.5279	0.011476
AR{1}	0.52337	0.075464	6.9354	4.0508e-12
SAR{12}	0.88377	0.041682	21.203	8.9963e-100
Variance	222.42	22.378	9.9393	2.8072e-23

Figura 17: Tabella riassuntiva del modello SAR(12)

Il modello è stato allenato sul dataset di training che presenta una buona stagionalità a ritardo 12, che però diminuisce fino a sparire nel dataset di test. Per questo motivo, nel momento in cui si valuta il modello sul dataset di test, non riesce a prevedere nel modo corretto la serie storica. Se si dovesse confrontare semplicemente AIC e BIC di questo modello con quello precedente, ci si accorgerebbe di un notevole miglioramento di performance. Poichè la valutazione di un modello deve essere fatta sul test set, si notano performance minori.

Modello ARIMA(2,0,2)

Per il motivo illustrato precedentemente (perdita di stagionalità della serie storica) si è deciso di applicare un modello che non tiene in considerazione la stagionalità dei dati, in quanto dal 2015 la serie storica non è più fortemente stagionale. Quest ultimo ottiene i risultati migliori per quanto riguarda la previsione delle emissioni della C0₂ negli ultimi 2 anni utilizzando il RMSE. Questa è la sua equazione analitica:

$$y_t = a_1 y_{t-1} + a_2 y_{t-2} + c_1 \varepsilon_{t-1} + c_2 \varepsilon_{t-2} + \varepsilon_t$$

Il risultato del modello è il seguente:

ARIMA(2,0,2) Model (Gaussian Distribution)				
Effective Sample Size: 115				
Number of Estimated Parameters: 6				
LogLikelihood: -529.839				
AIC: 1071.68				
BIC: 1088.15				
	Value	StandardError	TStatistic	PValue
Constant	422.57	15.782	26.775	6.323e-158
AR{1}	0.98688	0.033894	29.117	2.2123e-186
AR{2}	-0.94196	0.022871	-41.185	0
MA{1}	-0.7019	0.052808	-13.292	2.5932e-40
MA{2}	0.86977	0.050836	17.109	1.2654e-65
Variance	587.99	70.548	8.3346	7.7765e-17

Figura 18: Tabella riassuntiva del modello ARIMA(2,0,2)

Modello RegARIMA(2,0,2)

Il modello RegArima presenta una parte autoregressiva e una a media mobile di ordine 2. Inoltre, abbiamo utilizzato come regressori 3 armoniche formate dalla funzione coseno utili per modellare meglio la stagionalità. Il risultato sopracitato ha prodotto un RMSE di 26.70. L'equazione del modello stimato è la seguente:

$$y_t = a_1 y_{t-1} + a_2 y_{t-2} + c_1 \varepsilon_{t-1} + c_2 \varepsilon_{t-2} + \beta_1 * \cos(f1) + \beta_2 * \cos(f2) + \beta_3 * \cos(f3) + \varepsilon_t$$

Regression with ARMA(2,2) Error Model (Gaussian Distribution):				
	Value	StandardError	TStatistic	PValue
Intercept	442.92	2.7272	162.41	0
AR{1}	0.98356	0.052947	18.576	4.9857e-77
AR{2}	-0.96128	0.033078	-29.061	1.1123e-185
MA{1}	-0.71327	0.053964	-13.217	6.9626e-40
MA{2}	0.90587	0.056197	16.119	1.862e-58
Beta(1)	18.475	4.6889	3.9402	8.1404e-05
Beta(2)	-3.3657	77.879	-0.043217	0.96553
Beta(3)	10.119	2.1338	4.7421	2.1154e-06
Variance	386.17	47.689	8.0978	5.5976e-16

Figura 19: Tabella riassuntiva del modello RegARIMA(2,0,2).

Il risultati ottenuti dai modelli ARMA sul test set riescono a catturare abbastanza bene un evento esogeno come il lockdown dovuto a COVID-19, che determina una diminuzione improvvisa e brusca delle emissioni di CO₂.

Confronto tra i modelli e discussioni

I modelli sono stati tutti valutati utilizzando il test set. In particolar modo, il test set comprende un evento esogeno estremo ossia, l'abbassamento delle emissioni di C0₂ dovute al lockdown durante il Marzo-Aprile 2020.

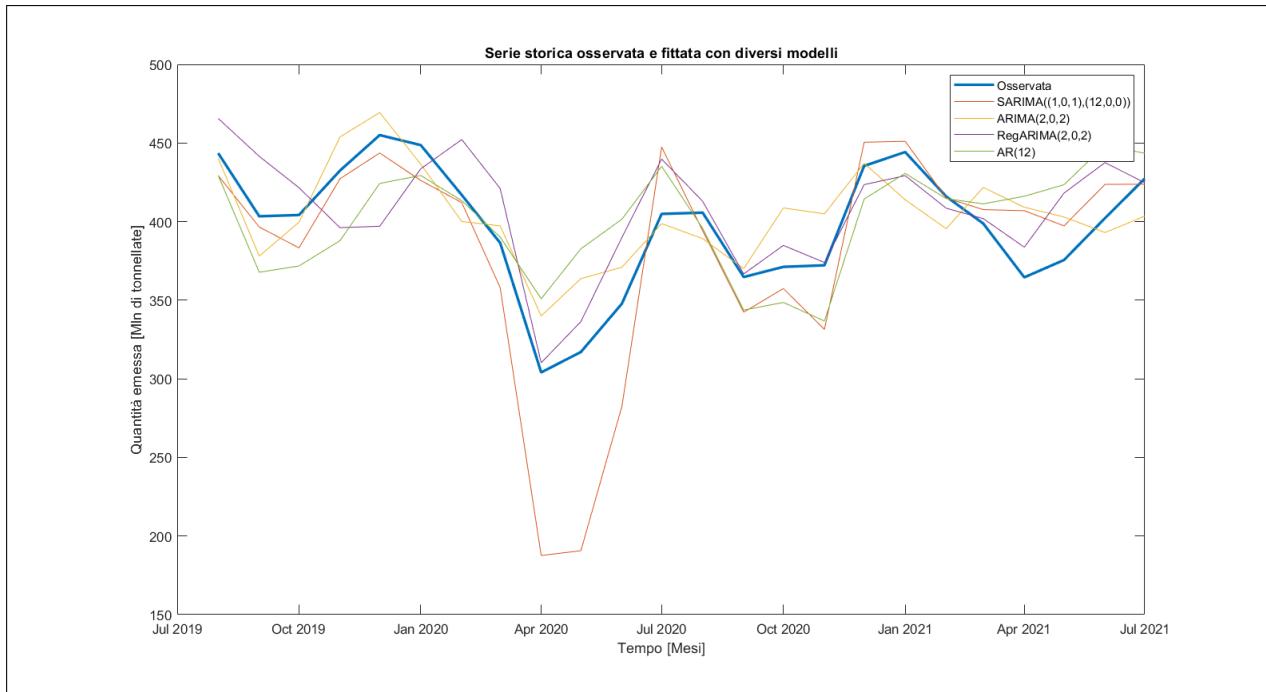


Figura 20: Tabella riassuntiva confronto modelli sul test set

I modelli sono stati valutati mediante un approccio quantitativo ossia valutando il Root Mean Squared Error (RMSE). I risultati sono riportati nella tabella a seguito:

	Valutazioni migliore fitting sul test set			
	AR (12)	SAR (12)	ARMA (2,0,2)	RegARIMA (2,0,2)
RMSE	33.31	42.52	24.13	26.70

Il modello maggiormente significativo, ossia ARIMA(2,0,2) ottiene i migliori risultati sul test set, nonostante sovrastimi leggermente le emissioni di CO₂ nel periodo del lockdown (Marzo-Aprile 2020).

Si è deciso di valutare come fossero le innovazioni del modello ARIMA(2,0,2) per capire se fossero $NID(0, \sigma^2)$ (processo white-noise). Nella sezione a seguito vengono riportati i risultati della modellazione dei residui.

Valutazione e Modellazione residui

Per prima cosa, sul test set si sono valutate:

- la normalità dei residui mediante il test di Jarque-Bera → normali
- la stazionarietà dei residui con il test di Dickey-Fuller → stazionari

Successivamente è stato necessario valutare l'autocorrelazione sui residui del modello. Ciò che si voleva ottenere erano dei residui che non presentassero autocorrelazione (processo white-noise). Questi sono i residui prima della modellazione:

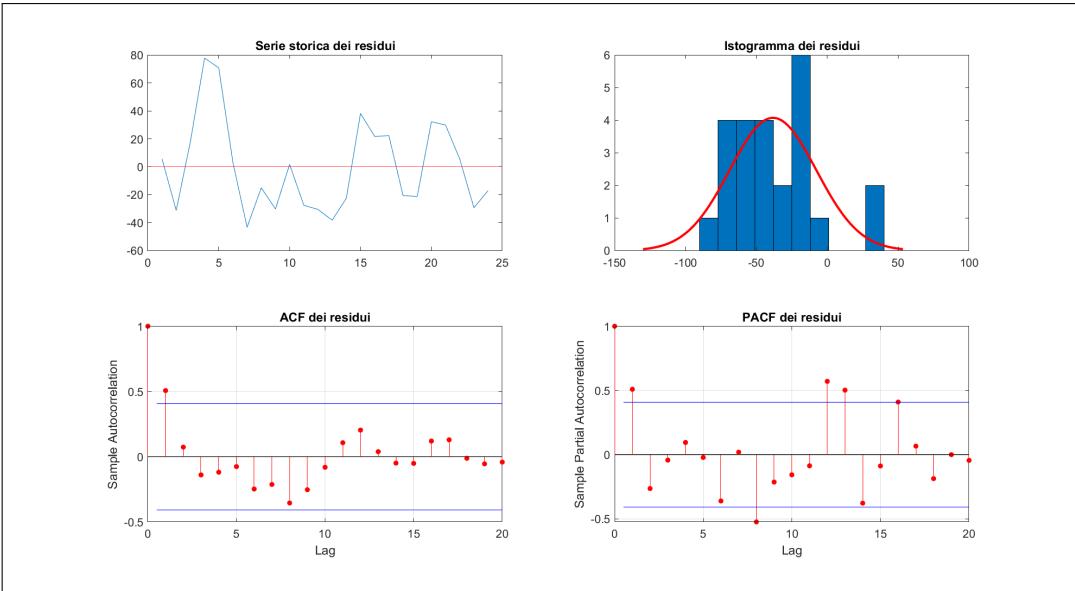


Figura 21: Tabella dei residui ARIMA prima della modellazione

Si nota una bassa autocorrelazione ACF e una buona autocorrelazione parziale PACF presente nei residui dell'ARIMA. In particolare i ritardi 1,8,12,13 dell'autocorrelazione parziale sono statisticamente significativi (test di Ljung-Box). Per questo motivo si è cercato di modellare i residui mediante un ARIMA (3,2,1). Il risultato ottenuto a seguito della modellazione è il seguente:

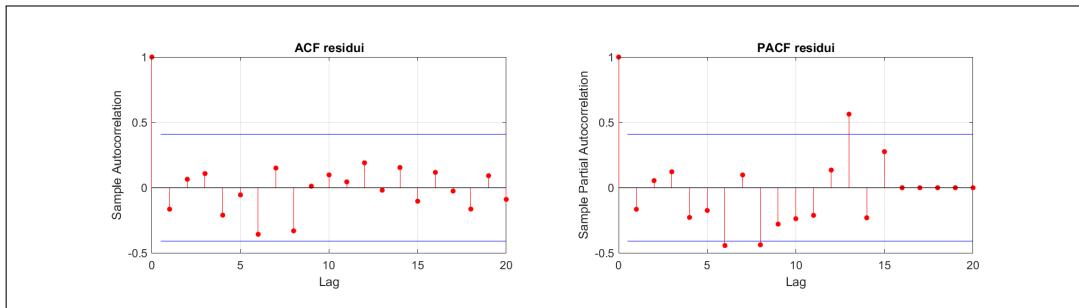


Figura 22: Tabella dei residui ARIMA dopo la modellazione

Dopo la modellazione non è più presente autocorrelazione ACF statisticamente significativa, mentre è stata ridotta notevolmente l'autocorrelazione parziale PACF, ma permane ancora una bassa autocorrelazione ai ritardi 8 e 13 nonostante non sia statisticamente significativa (verificato con test Ljung-Box).

4 Relazione tra emissioni di CO₂ e HDD

Stima delle emissioni di CO₂ basata su indici climatici

Contesto

Il periodo di interesse dell'analisi copre il periodo Gennaio 2010-Luglio 2021 e si considerano le seguenti variabili di studio:

- Sull'asse delle X, HDD [gradi giorno];
- Sull'asse delle Y, Quantità totale di CO₂ emessa.

Un grado giorno di riscaldamento (HDD - Heating Degree Days) è una misura progettata per quantificare la domanda di energia necessaria per riscaldare un edificio. È il numero di gradi per cui la temperatura media giornaliera è inferiore a 65° Fahrenheit (18° Celsius), che corrisponde alla temperatura al di sotto della quale gli edifici devono essere riscaldati. Numericamente sono calcolati come la somma cumulativa della sola differenza positiva tra la temperatura interna di base e la temperatura media esterna.

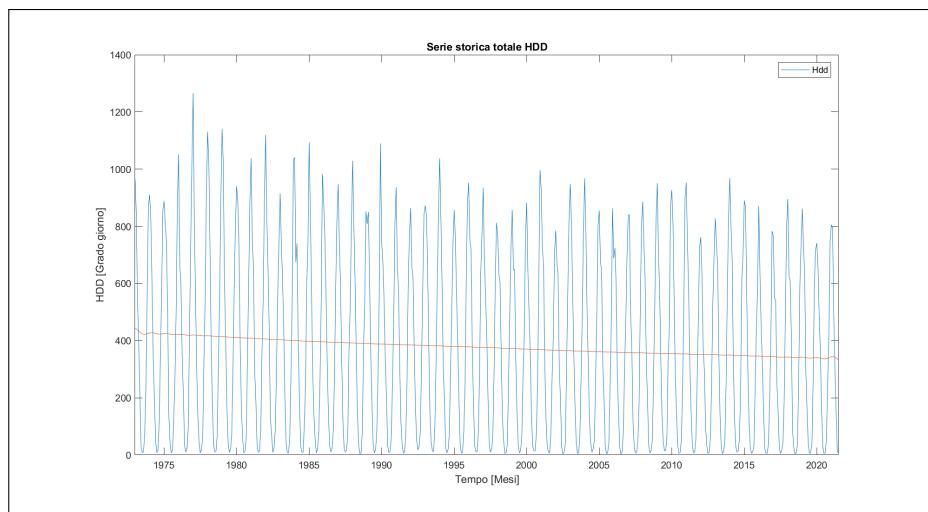


Figura 23: Serie storica dell'HDD

Come mostrato in figura 23, il trend della serie storica dell'HDD è decrescente, ciò è una conferma del problema del surriscaldamento globale: infatti la temperatura media globale sta aumentando, fa sempre meno freddo e quindi la quantità di energia necessaria per riscaldare gli ambienti è sempre in diminuzione.

Analisi delle emissioni con regressione statica utilizzando HDD

La correlazione tra le emissioni totali di CO₂ e HDD porta ad un risultato di 0.4486 mentre la correlazione con gli HDD² risulta essere di 0.5951, più forte rispetto al caso precedente.

Nel grafico seguente è rappresentata la relazione tra HDD e le emissioni di CO₂; come è possibile notare, una regressione lineare non descrive bene l'andamento delle emissioni. Al contrario il modello quadratico risulta essere molto più interessante in quanto segue l'andamento di quest'ultime.

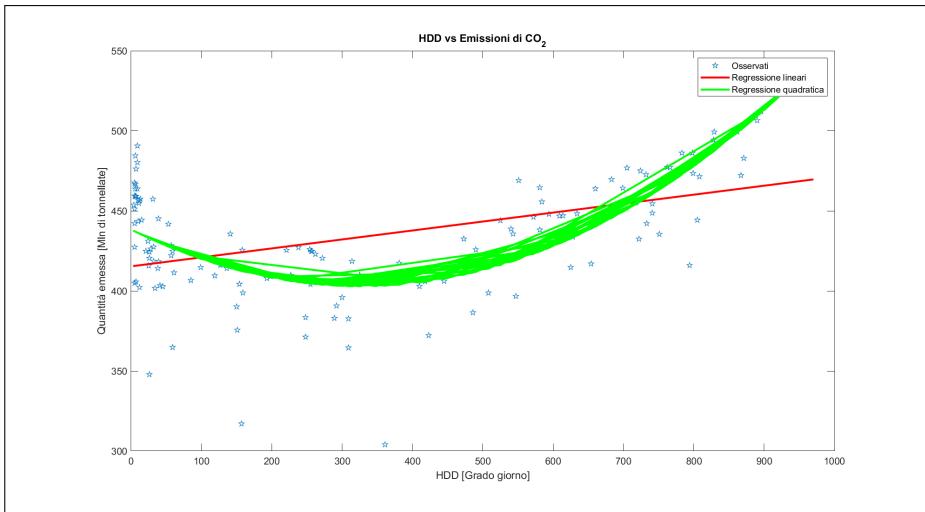


Figura 24: Confronto emissioni-HDD con andamento lineare e quadratico

Per l'analisi in questione sono stati utilizzati due modelli: il primo con un solo regressore (HDD) mentre il secondo con due regressori (HDD e HDD^2). Per quanto riguarda le emissioni fittate lineari si nota che R^2 non è per nulla significativo (0.201). Ciò è ben visibile anche dal grafico in figura 25 in quanto non si considerano i picchi di Luglio; infatti presenta un andamento simile ad una sinusoide, molto regolare. Nel secondo modello cresce notevolmente l' R^2 (0.572), in questo caso rispetto al precedente è possibile notare che i picchi di Luglio vengono considerati, anche se sottostimati. Inoltre, è possibile vedere una buona relazione in corrispondenza dei mesi Gennaio (picchi alti) e dei mesi di Aprile (picchi bassi). Tuttavia, entrambi i modelli non riescono a valutare il fenomeno esogeno COVID-19 di Marzo/Aprile. L'andamento delle emissioni stimate non subisce una netta variazione tra Aprile 2019 e Aprile 2020 in entrambi i modelli. Il motivi sono duplici:

- le emissioni sono calate drasticamente durante il lockdown mentre l'indice HDD non si è modificato rimanendo pressoché costante;
- l'utilizzo di una regressione statica non permette di cogliere andamenti tempo-varianti (COVID-19).

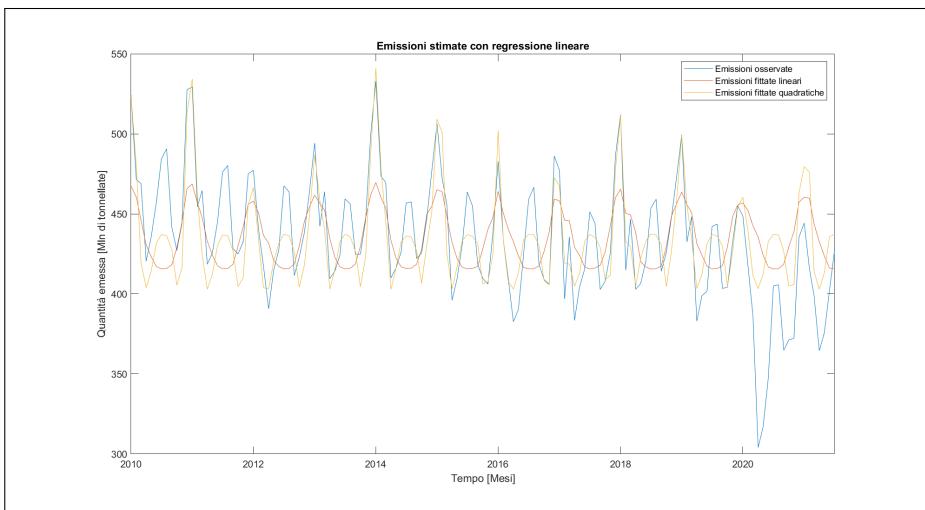


Figura 25: Confronto utilizzando i modelli di regressione statici

Analisi delle emissioni con regressione dinamica (modello state-space) utilizzando HDD

Un modello di state-space è una forma molto generale di rappresentare sistemi dinamici, particolarmente adatta a fare inferenza su componenti non osservabili. Nel nostro caso, sono stati creati due modelli con l'obiettivo di valutare se ci fossero dei miglioramenti sulla stima nel periodo di COVID-19. L'obiettivo è minimizzare l'AIC e il BIC, due parametri che forniscono una misura della qualità del modello ottenuta simulando la situazione in cui il modello viene testato su un diverso set di dati. Quindi più bassi sono i valori che otteniamo più il modello è simile alla realtà.

1 - Modello state-space con alpha costante e beta variabile: In questo primo modello, l'intercetta (alpha) è costante mentre, il coefficiente angolare (beta) è tempo-variante. Il primo passaggio è il filtraggio dei dati e successivamente si applica lo smooting degli stati. La funzione smooth(y) uniforma i dati di risposta nel vettore colonna y utilizzando un filtro a media mobile. In questo caso si ottengo un valore AIC=1420.31 e un BIC= 1426.17. Analizzando visivamente il grafico (figura 26) si nota che nessuno dei due si comporta correttamente. Tuttavia, apparentemente sembrano meglio le emissioni filtrate rispetto a quelle smussate, anche se non riescono a valutare il picco del lockdown. Verificando poi i valori degli R^2 , si verifica l'ipotesi fatta in precedenza. Infatti, R^2 filtered=0.3958 mentre R^2 smoothed= 0.3576.

2 - Modello state-space con alpha variabile e beta variabile: In questo secondo modello, sia l'intercetta (alpha) che il coefficiente angolare (beta) sono tempo-varianti. In questo caso si è applicato il medesimo procedimento del modello precedente, ottenendo un valore AIC=1399.13 e un BIC= 1407.93. Analizzando visivamente il grafico si nota che anche in questo caso apparentemente sembrano meglio le emissioni filtrate rispetto a quelle smussate, tuttavia, rispetto al caso precedente, il picco del lockdown viene in parte considerato. Ciò è giustificato anche dai valori degli R^2 . Infatti, analizzandoli, si verifica l'ipotesi fatta in precedenza ovvero, R^2 filtered=0.5733 mentre R^2 smoothed= 0.5192

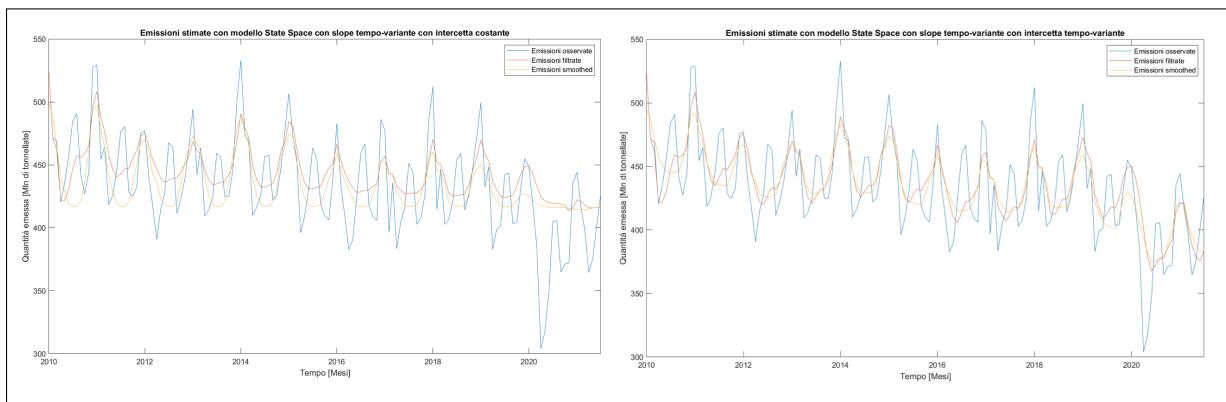


Figura 26: A sx: modello 1 di regressione dinamica / A dx: modello 2 di regressione dinamica

Confronto migliore modello statico e dinamico

Si è rappresentato in figura 27 il confronto tra il miglior modello statico (quello quadratico) e il miglior modello dinamico (quello con alfa e beta variabili). I due modelli presentano R^2 molto simili ma si comportano diversamente: mentre il modello statico presenta un andamento abbastanza regolare, stimando correttamente tutti

i picchi ad eccezione del lockdown, il modello dinamico sottostima leggermente i picchi ma riesce a individuare il picco del lockdown e anche quello successivo. In conclusione si può dire che, sia il modello statico che quello dinamico, che si servono dell'HDD come previsore, non sono molto significativi. Ciò avviene perché le emissioni sono state previste basandosi su un fattore climatico e nonostante esse stiano diminuendo negli ultimi anni (soprattutto durante il lockdown), l'HDD invece resta abbastanza costante. Nel materiale allegato al progetto è stato provato un modello ibrido, cioè statico per i primi 9 anni e dinamico per gli ultimi 2 anni (R^2 migliora a 0.68).

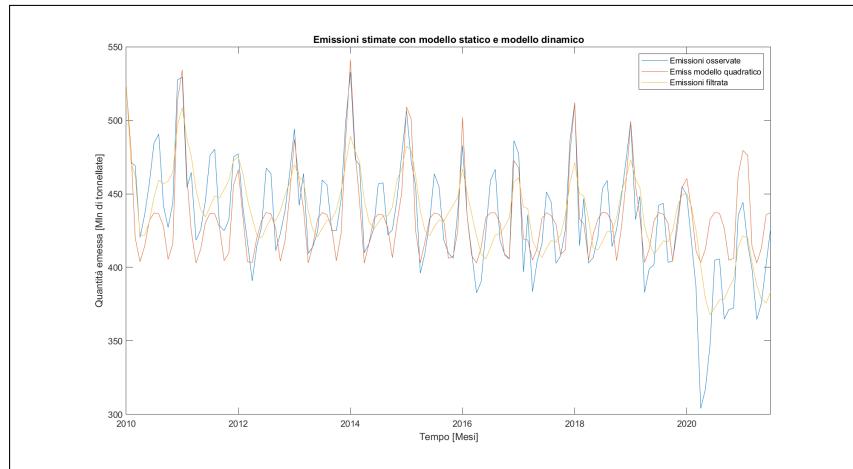


Figura 27: Confronto migliore modello statico (quadratico) e dinamico (alfa e beta variabili)

Analisi e modellazione residui del modello HDD²

Si è svolta poi l'analisi sui residui del modello HDD² che presentava un elevato R^2 . Come è possibile vedere dalla figura 28, la distribuzione è circa centrata in zero e la maggior parte si trova sotto la campana con asimmetria a destra. Inoltre, applicando gli opportuni test, si nota che i residui non sono normali e tramite il test di Engle, è possibile affermare che essi sono eteroschedastici.

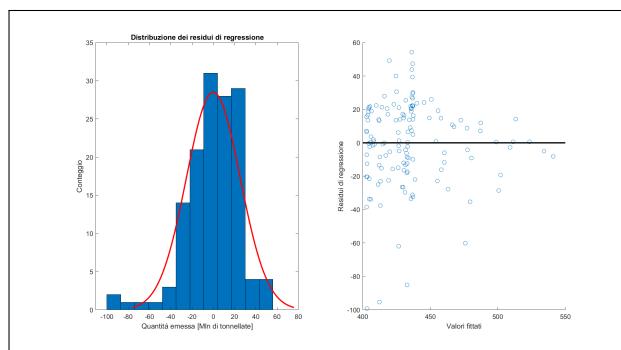


Figura 28: Residui del modello di regressione quadratico

Si è poi verificata l'autocorrelazione parziale e totale dei residui e dei residui al quadrato, visibile in figura 29.

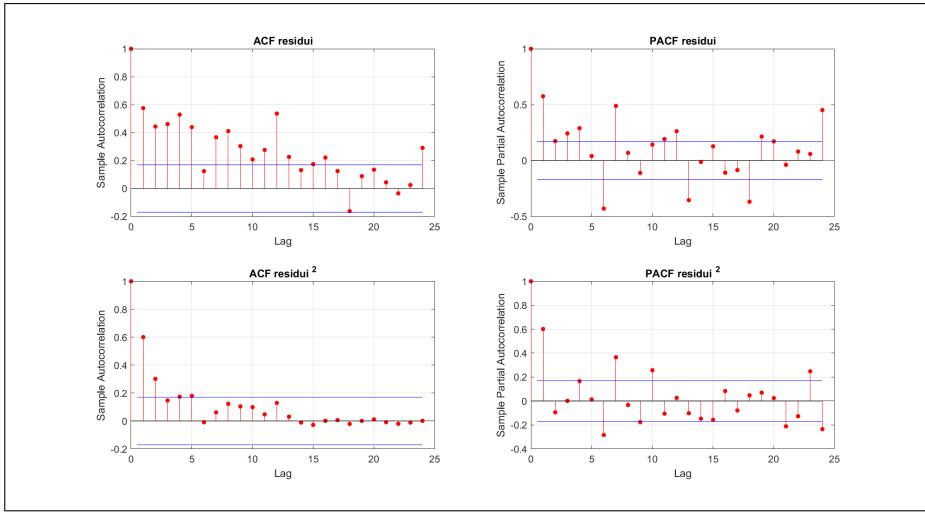


Figura 29: Autocorrelazione del modello di regressione statico

Pre-whitening con ARMA(3,0,5)

Si è applicato un modello ARMA con l’obiettivo di eliminare l’autocorrelazione totale e parziale. E’ possibile notare in figura 30 che, in seguito all’applicazione del modello, sia ACF che PACF sono diminuite. Tuttavia, in entrambi i casi non tutti i valori rientrano nelle bande, ciò è però giustificato dal fatto che, essendo un dataset con dati reali, non è sempre possibile creare un modello perfetto.

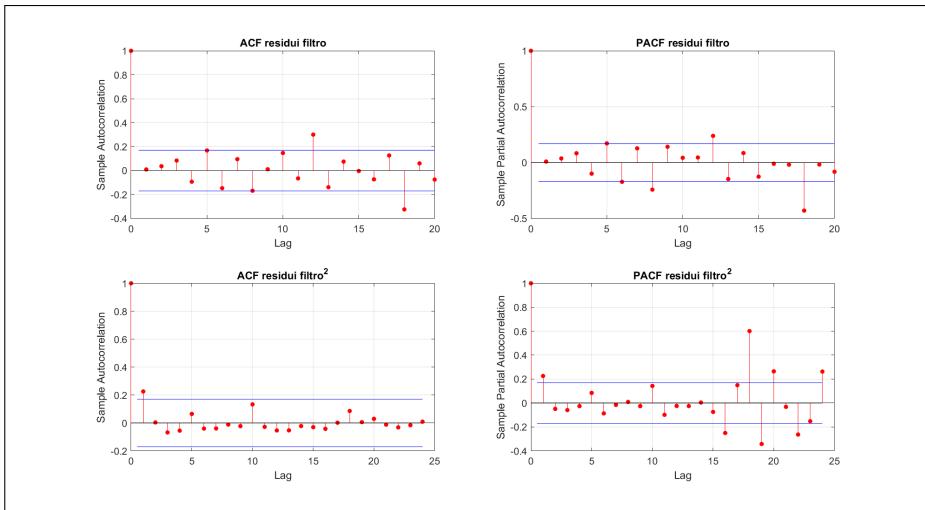


Figura 30: Autocorrelazione del modello di regressione statico dopo ARMA

Studio della variabilità residua con GARCH (0.2)

Come analisi aggiuntiva si è pensato di provare a modellare i residui al quadrato tramite GARCH (0.2), che presenta solo parte autoregressiva. Il modello GARCH è un modello autoregressivo generalizzato che cattura i raggruppamenti di volatilità dei rendimenti attraverso la varianza condizionale. Viene utilizzato per la sua capacità di prevedere la volatilità a breve e medio termine. Il modello GARCH trova la volatilità media attraverso un’autoregressione che dipende dalla somma degli shock ritardati e dalla somma delle varianze ritardate.

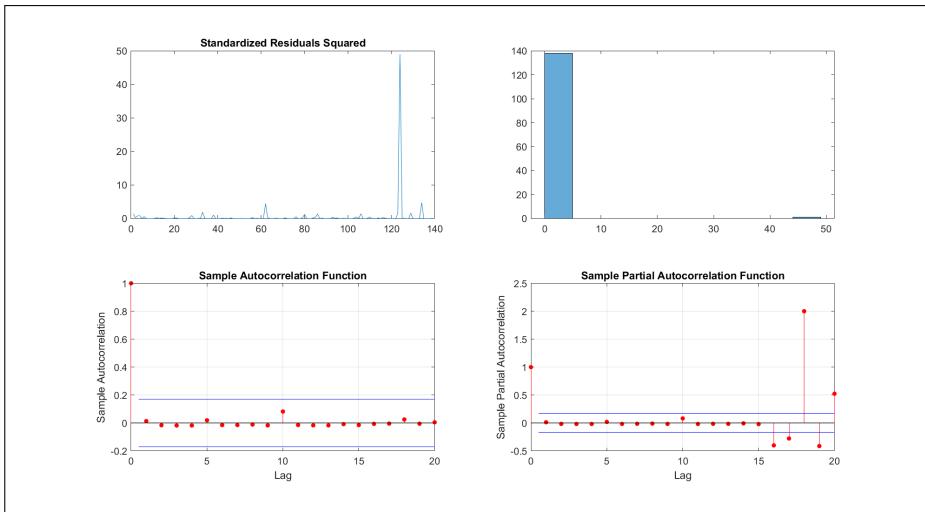


Figura 31: Residui standardizzati al quadrato dopo GARCH

I residui al quadrato presentano ancora autocorrelazione parziale ma non importante come nel caso precedente.

5 Stima delle anomalie delle temperature globali basata sulle emissioni di CO₂

Il riscaldamento globale indica il mutamento del clima terrestre sviluppatosi a partire dalla fine del XIX secolo e tuttora in corso, caratterizzato in generale dall'aumento della temperatura media globale e da fenomeni atmosferici a esso associati. Le cause predominanti sono da ricercare principalmente nell'attività umana, in ragione delle emissioni nell'atmosfera terrestre di crescenti quantità di gas serra.

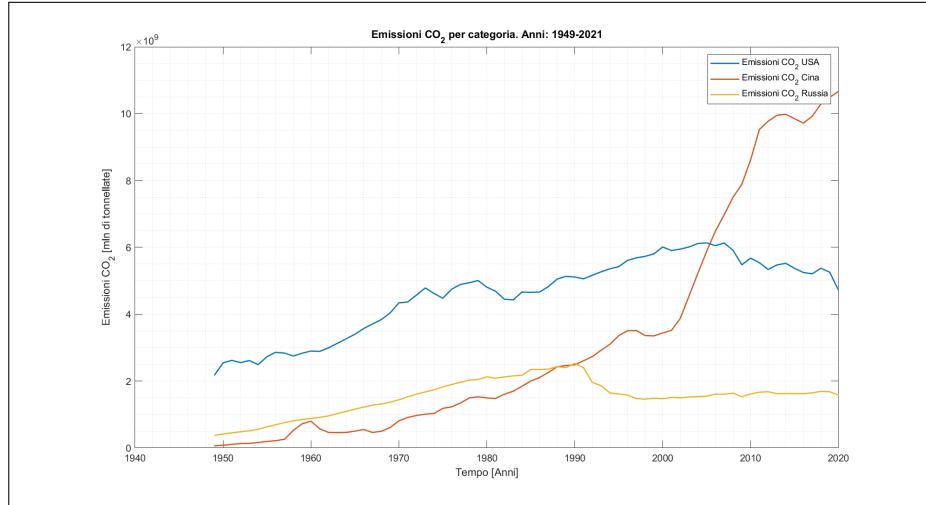


Figura 32: Confronto emissioni di CO₂ tra USA, Cina e Russia

Regressione lineare

Da questo momento si è considerato il dataset con i dati annuali. Applicando una regressione lineare semplice con un solo predittore (le emissioni di CO₂ degli USA) si ottiene un R² abbastanza significativo pari a 0.548. In figura 33 è stato rappresentato il confronto tra le anomalie reali, provenienti dal dataset e quelle stimate.

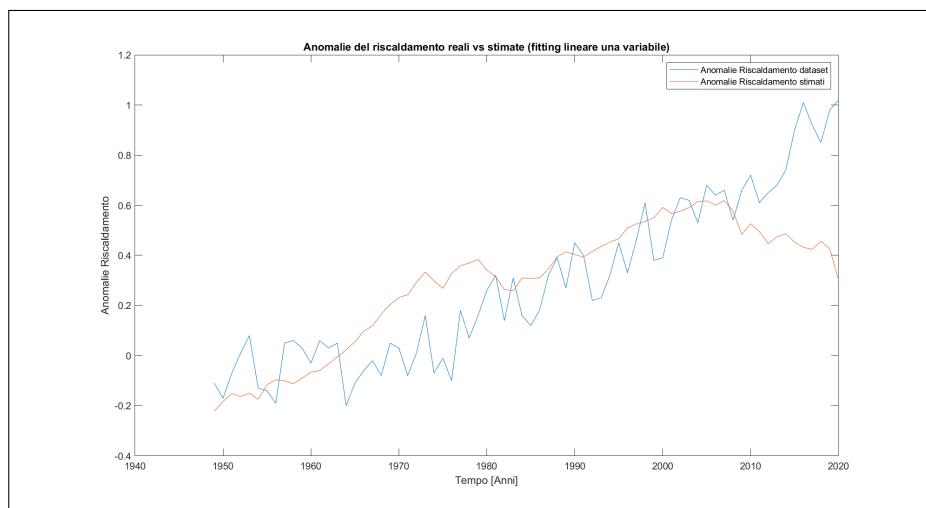


Figura 33: Confronto anomalie del riscaldamento (dati reali vs dati stimati)

Il risultato non è molto soddisfacente come è possibile notare dal grafico dove nell'ultimo periodo la stima si discosta parecchio dalla realtà. Infatti, a partire dal 2007, gli USA hanno adottato una serie di politiche volte a ridurre le emissioni di CO₂ mentre le anomalie sul riscaldamento sono attualmente in continuo aumento.

Regressione Multipla

Vista la scarsa significatività del precedente modello è stata realizzata una regressione multipla utilizzando più predittori. Quello che ci si aspetta è un modello ovviamente migliore in quanto si prendendo in considerazione tre grandi potenze mondiali quali USA, Cina e Russia, responsabili della maggior parte delle emissioni di CO₂, per stimare le anomalie delle temperature globali. Analizzando i p-value, il più significativo risulta essere quello della Cina, le cui emissioni risultano molto correlate con le anomalie. Un R²=0.885 indica la maggiore significatività del modello rispetto a quello stimato con la regressione semplice.

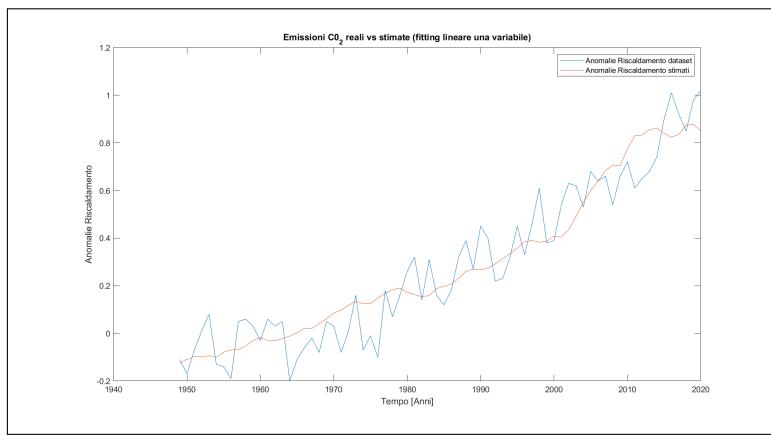


Figura 34: Confronto anomalie del riscaldamento (dati reali vs dati stimati)

Analisi dei residui del miglior modello

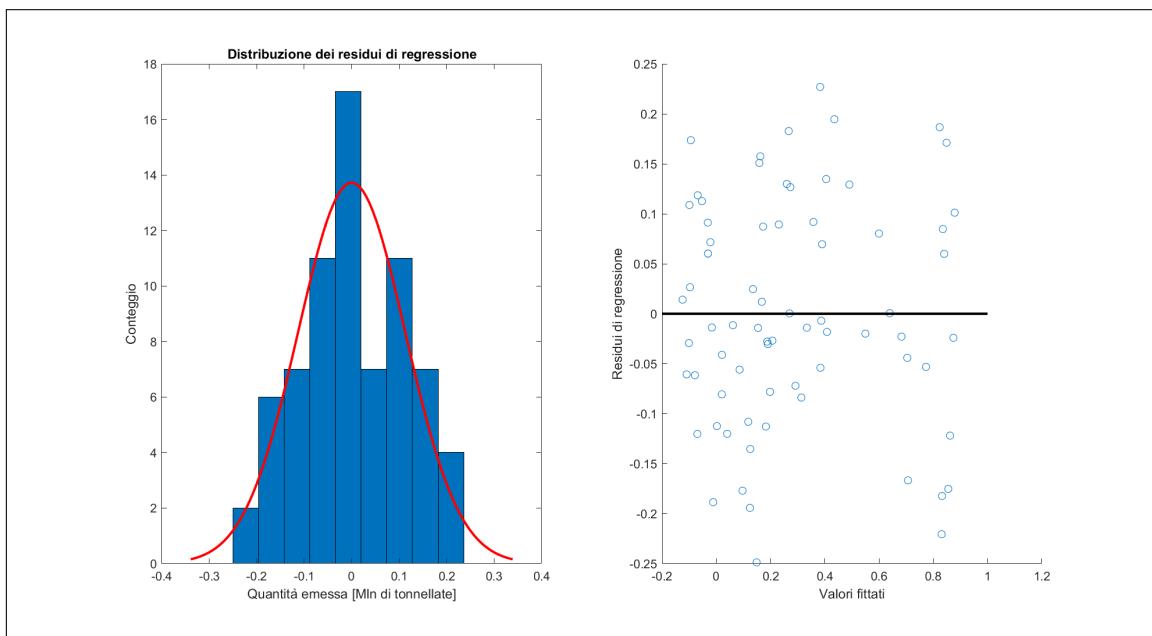


Figura 35: Residui del modello di regressione multipla

Il test di Bera-Jarque con livelli di significatività dell'1% e del 5% indica che i residui sono normali, ipotesi confermata anche dal successivo test di Lilliefors.

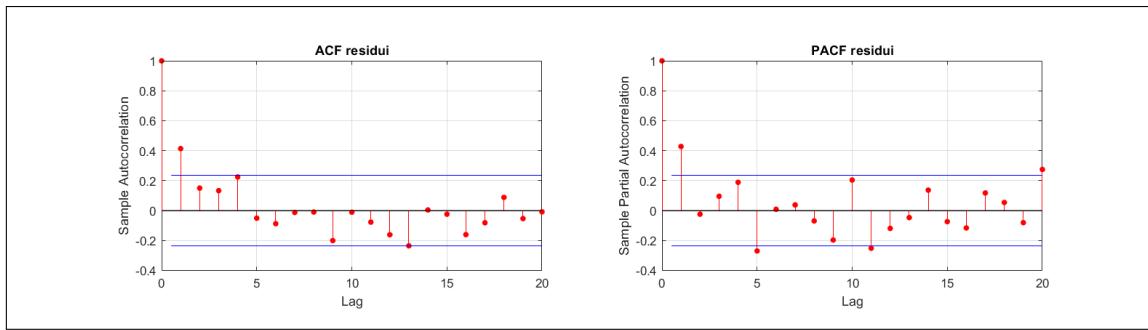


Figura 36: Autocorrelazione dei residui del modello di regressione multipla

I residui risultano poco autocorrelati.

5.1 Modellazione residui con ARIMA

Per modellare meglio i residui della regressione si è applicato un ARIMA(1,0,0), ma i risultati del correlogramma in figura 37 mostrano ancora autocorrelazioni totali a ritardi 4,9,13 e parziali a 4,9 e 20:

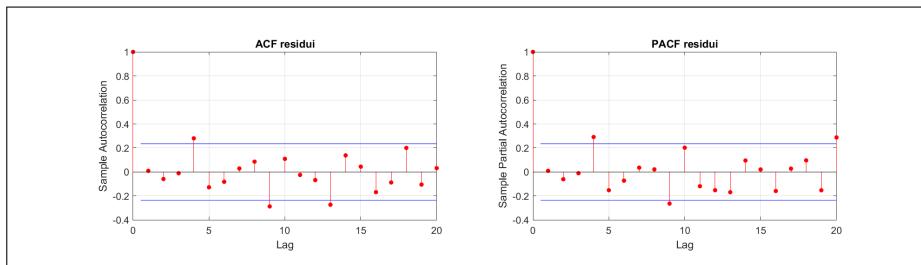


Figura 37: Correlogramma ARIMA(1,0,0)

Per eliminare l'autocorrelazione totale e parziale si è deciso di applicare un ARIMA(2,0,1):

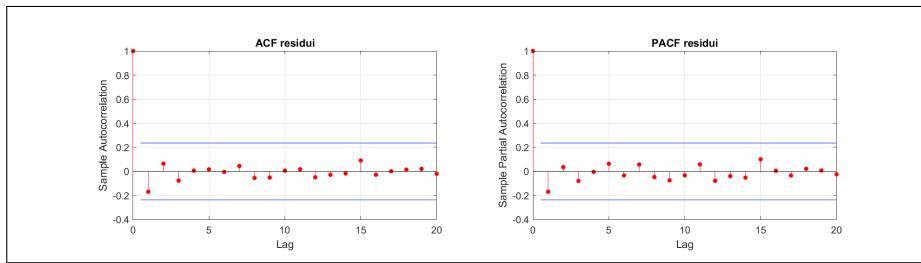


Figura 38: Correlogramma ARIMA(2,0,1)

Quest'ultimo modello rende i residui della regressione non autocorrelati.

6 Discussione e Conclusioni

Nel presente elaborato è stato mostrato come i fattori produttivi influenzano le emissioni di CO₂, mettendo in luce in maniera inequivocabile la relazione di causalità tra i due; al crescere di determinati consumi o produzioni di combustibili fossili, particolarmente inquinanti, crescevano le emissioni.

Successivamente è stata valutata la possibilità di modellare l'andamento delle emissioni di CO₂ con l'ausilio di modelli autoregressivi ossia, basati solamente sull'andamento precedente. Quello che si è potuto valutare è stata una migliore accuratezza dei modelli regressivi, i quali utilizzano come variabili predittive fattori determinanti come i consumi di combustibili.

Un'altra relazione indagata è stata il rapporto tra Emissioni di CO₂ e l'indice HDD: mediante questo indice climatico risulta molto difficile stimare le emissioni di CO₂ di un paese come gli Stati Uniti.

Infine, un risultato molto interessante è stato identificato mediante l'utilizzo di dati annuali: è stato valutato in che modo i 3 paesi che emettono più CO₂ al mondo stanno influenzando il riscaldamento climatico presente sul pianeta. In particolare, risulta molto evidente come le emissioni in Cina risultino il fattore maggiormente predittivo per stimare le anomalie di temperatura che si registrano sul pianeta Terra. Gli USA risultano leggermente meno predittivi poichè negli ultimi 20 anni sono state intraprese campagne di sensibilizzazione sul tema del riscaldamento climatico. Nonostante ciò, le anomalie delle temperature stanno continuando a crescere, probabilmente poichè il clima è un fenomeno molto più complesso che non dipende solamente dalle emissioni di CO₂ dei paesi più industrializzati.

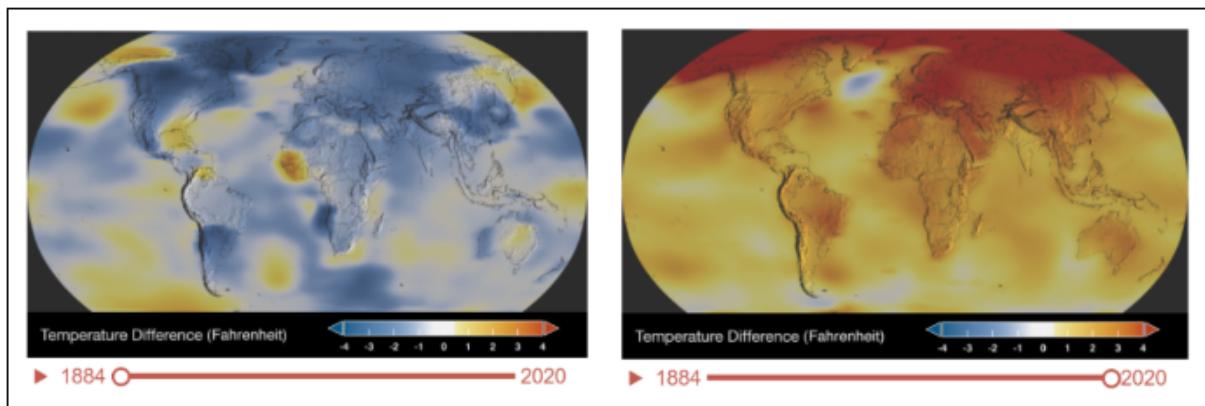


Figura 39: Confronto tra le temperature presenti nel 1884 e quelle presenti nel 2020.

Questa continua crescita dovrebbe essere un segnale d'allarme poichè dal 1884 al 2020 è stato registrato un aumento delle temperature sul pianeta terra molto significativo (nas (2021)). Basti pensare che la temperatura superficiale globale è aumentata di 0,2°C per decennio negli ultimi 30 anni (Hansen et al. (2006)). Lo studio e la ricerca di fonti di energia rinnovabili sono la strada da seguire per un futuro più sostenibile e pulito.

Fonti dataset

Le fonti utilizzate per la composizione del dataset sono le seguenti:

- I dati denominati “Table_NUMBER” sono stati presi dalla United States Energy Information Administration (EIA): <https://www.eia.gov/totalenergy/data/annual/>
- I dati denominati “Tab_NUMBER” sono stati ricavati dalla piattaforma online “Our World In Data”. Si tratta di un sito di pubblicazione scientifica appartenente alla categoria di Editoria digitale: <https://ourworldindata.org/co2/country/china?country=CHN USA RUS>
- I dati annuali sulle anomalie del riscaldamento globale sono presenti a questo sito: <https://climate.nasa.gov/vital-signs/global-temperature/>
- I dati della vendita automobili sono presenti al sito: <https://www.goodcarbadcar.net/usa-auto-industry-total-sales-figures/> che fa riferimento al sito USA Bureau of Economic Analysis.

Per visualizzare i dati originali le allego il link al repository Git-Hub utilizzato per tenere traccia di tutte le modifiche allo script e ai metadati: https://github.com/M-ballabio1/Energy_Statistics_project

Riferimenti bibliografici

- 2021, Global surface temperature, NASA. <https://climate.nasa.gov/vital-signs/global-temperature/>
- Dagum, E. 2001, Analisi delle serie storiche: modellistica, previsione e scomposizione, Collana di Statistica e Probabilità Applicata (Springer). <https://books.google.it/books?id=8GnIHdjsDJsC>
- Hansen, J., Sato, M., Ruedy, R., et al. 2006, Proceedings of the National Academy of Sciences, 103, 14288
- Shindell, D., Ru, M., Zhang, Y., et al. 2021, Proceedings of the National Academy of Sciences, 118