Università degli Studi di Padova
Dipartimento di Scienze Statistiche
Corso di Laurea Triennale in

Statistica per l'Economia e l'Impresa

RELAZIONE FINALE

SERIE STORICHE IRREGOLARI: UN'ANALISI CON I DATI PALEOCLIMATICI

Relatore Prof.ssa Luisa Bisaglia
Dipartimento di Scienze Statistiche

Laureando: Alex De Meneghi
Matricola N 1166608

Anno Accademico 2019/2020

Indice

1 Introduzione

L'analisi di serie storiche multivariate si occupa dell'interpretazione e dello studio di serie di dati che vengono osservati regolarmente, alla stessa frequenza e contemporaneamente. Quelle che fanno parte della nostra analisi, invece, sono osservate a frequenze differenti e forniscono un esempio in cui i dati non sono osservati contemporaneamente. Queste serie storiche, come citato da Miller (2019), vengono chiamate *'serie storiche a frequenza mista'*, e comprendono sia il caso in cui la frequenza sia regolare ma differente, sia quello in cui una o più serie cambiano la frequenza delle osservazioni durante il periodo di osservazione. Di queste serie storiche se ne sono occupati, tra gli altri, Busetti e Taylor (2005), Ghysels e Miller (2014) e Ghysels e Miller (2015) che le hanno analizzate e hanno applicato alcuni metodi per la loro trattazione e per risolvere il problema dell'*irregolarità* delle serie di cui ci occupiamo. Alcuni di questi metodi li approfondiremo nel seguito.

Pensando ad un esempio di questa tipologia di serie in ambito finanziario, uno dei più importanti è quello riguardante i prezzi azionari giornalieri negoziati nei mercati in differenti fusi orari (non contemporaneità dei dati), non raccolti durante il weekend (dati spaziati in modo irregolare), e soprattutto provenienti da diverse nazioni del mondo, dove le vacanze sono differenti in base a dove si trova il paese (dati irregolari), Miller (2019). Tra gli esempi di serie irregolari c'è anche quello riguardante i dati paleoclimatici, oggetto della nostra analisi.

Per chiarire ciò che analizzeremo, diamo innanzitutto una definizione del termine 'paleoclimatologia'. La paleoclimatologia può essere definita come *"La disciplina che studia le condizioni geografiche della Terra nei vari periodi della storia geologica (paleogeografia), in particolare la distribuzione e la successione dei climi"*, Zingarelli (2010). I dati che analizzeremo, relativi a questa disciplina e che risulteranno molto importanti, sono le serie della concentrazione di anidride carbonica in ppmv (*parts per million by volume*) e quelle della temperatura misurate in gradi Celsius. Se si cercasse di trovare una relazione tra queste serie particolari si potrebbe discutere su dei possibili effetti che questi due fenomeni congiuntamente creeranno in futuro.

I dati che vengono analizzati vengono ricavati dalla perforazione dei ghiacci, soprattutto nelle zone come l'Antartide o la Groenlandia. Nella sezione relativa all'applicazione ai dati paleoclimatici è approfondito il processo

di raccolta dei dati.

Il metodo che viene principalmente utilizzato in ambito statistico per verificare se è presente correlazione tra serie storiche che hanno trend stocastici, i quali si muovono in modo simile nel lungo periodo tanto da sembrare gli stessi per entrambe le serie, è l'analisi della cointegrazione. Questo strumento è stato applicato molte volte ai dati paleoclimatici, soprattutto da Kaufmann e Juselius (2013), i quali hanno stimato un CVAR (vedi capitolo 3) applicato a dei dati relativi al clima per verificare quali sono i meccanismi che ci sono all'interno dei cicli glaciali durante l'ultima era geologica. Da questa analisi hanno scoperto che c'è cointegrazione tra due serie particolari tra tutte quelle analizzate: quella della temperatura e quella della concentrazione di anidride carbonica nell'aria, le quali verranno discusse nella sezione dell'applicazione ai dati paleoclimatici.

L'articolo è strutturato nel modo seguente. Il capitolo 2 introduce brevemente il concetto di cointegrazione tra due o più serie storiche e discute dei test più comuni che vengono applicati per verificarne la presenza. Il capitolo 3 contiene la definizione e la rappresentazione dei VAR (Vector Autoregressive Model), i quali sono tra i modelli più applicati in ambito statistico, e presenta un esempio per vedere com'è costituito e come va applicato questo tipo di modello. Il capitolo 4 si occupa, invece, dei CVAR (Cointegrated Vector Autoregressive Model) cercando di darne una definizione e una successiva rappresentazione matematica per l'analisi di cointegrazione. Il capitolo 5 introduce il concetto di serie storica irregolare e discute vari metodi che possono essere utilizzati per trattare questa particolare e insolita tipologia di serie. Per concludere è presente il capitolo 6, il più importante dell'articolo, che riguarda la vera e propria applicazione ai dati paleoclimatici e contiene una serie di simulazioni di Monte Carlo per discutere della potenza e della dimensione empirica dei test che vengono applicati.

2 La cointegrazione

2.1 Definizione e rappresentazione

Il concetto di cointegrazione è molto importante ai nostri fini, in quanto utile per la definizione di modelli VAR (Vector Autoregressive Model, i quali verranno spiegati nella sezione successiva), nei quali è importante valutare la non stazionarietà della maggior parte delle variabili che vengono analizzate. Esso fu introdotto, come citato da Watson (1994), in una serie di articoli di Granger e Weiss (1983) e di Engle e Granger (1987). Questi articoli sviluppano dei metodi per la trattazione e l'analisi di relazioni economiche, sia a breve che a lungo termine.

Quando in ambito statistico si utilizzano dei modelli di cointegrazione si fa riferimento all'*integrazione* tra variabili (che verrà spiegata in questa sezione) e, come citato da Watson (1994), le statistiche costruite attraverso delle variabili integrate spesso non si comportano in modo standard, facendo sorgere dei problemi relativi, ad esempio, alla presenza di una radice unitaria, la quale non è facile da trattare con questa tipologia di variabili.

Per dare una definizione formale di 'cointegrazione' prendiamo quella citata da Peracchi (2012) che dice che *"Il fenomeno della cointegrazione si verifica nel caso in cui due o più serie temporali con trend stocastici si muovono congiuntamente in modo simile nel lungo periodo, tanto che sembrano possedere lo stesso trend"*. Quando parliamo di trend dobbiamo prestare attenzione nel distinguere tra:

- trend deterministico
- trend stocastico (come nel nostro caso)

Il primo è descritto da una funzione deterministica (e quindi non aleatoria) del tempo f(t) ed è "prevedibile" una volta che i coefficienti che lo specificano sono tutti noti. Il secondo, invece, è caratterizzato dal fatto che la componente di fondo varia nel tempo in maniera aleatoria (casuale), e quindi non è completamente prevedibile come quello deterministico.

Per spiegare la cointegrazione dobbiamo prima definire il concetto di integrazione di ordine 1; si veda, per esempio, Peracchi (2012). Una serie storica temporale X_t si dice integrata di ordine 1 se non è stazionaria, ma la serie che si ottiene differenziandola, ossia $\Delta X_t = X_t - X_{t-1}$ è stazionaria. Due serie storiche X_t e Y_t integrate di ordine 1 si dicono cointegrate se

esiste un coefficiente δ, detto anche coefficiente di cointegrazione, tale che la differenza $Y_t - \delta X_t$ sia una serie storica stazionaria. Perciò, ad esempio, la serie storica Z_t risulta essere cointegrata di ordine $(1, 1)$, ossia $Z_t \sim CI(1, 1)$, se tutte le sue componenti sono integrate di ordine 1, $I(1)$, e se esiste un vettore α diverso da zero tale che la serie ottenuta dalla moltiplicazione vettoriale $\alpha' Z_t$ sia integrata di ordine 0. In questo caso il vettore α viene chiamato 'vettore di cointegrazione'.

Quando due serie storiche X_t e Y_t non risultano essere cointegrate possono essere modellate congiuntamente attraverso un modello VAR alle differenze prime, ossia alle serie ΔX_t e ΔY_t. Questa tipologia di modellazione, approfondita nella sezione successiva, è un'estensione dei modelli autoregressivi $AR(p)$, i quali vengono descritti dall'equazione

$$Y_t = \phi_0 + \phi_1 Y_{t-1} + \cdots + \phi_p Y_{t-p} + \epsilon_t \tag{1}$$

dove Y_t è il processo autoregressivo di ordine p, ϕ_i per $i = 1, \ldots, p$ sono parametri costanti e $\{\epsilon_t\}$ è un processo white noise con media nulla e varianza costante σ_ϵ^2.

2.2 I test di cointegrazione

Esistono molteplici test per verificare se è presente cointegrazione tra due o più serie storiche. I tre più conosciuti, Miller (2019), sono:

- test della traccia, il quale è una delle due tipologie del test di Johansen (la seconda è il test del massimo autovalore) e verifica l'ipotesi nulla che il numero di vettori di cointegrazione sia inferiore al numero delle serie su cui verifichiamo la presenza di cointegrazione (questo test sarà utilizzato nella parte di applicazione ai dati paleoclimatici)

- test di Dickey-Fuller basato sui residui, che verifica la presenza di una radice unitaria nei residui di una relazione di cointegrazione tra diverse serie storiche

- t-test che, attraverso il calcolo di una statistica test (vedi $\hat{\tau}_M$ nel paragrafo successivo) e il successivo confronto con un valore nominale dello 0,05, ci porta a rifiutare o non rifiutare l'ipotesi di presenza di cointegrazione nell'analisi.

Per quanto riguarda le distribuzioni delle statistiche test sotto l'ipotesi nulla di non cointegrazione (contro l'alternativa che ne sia presente una qualche forma), utilizzando serie analizzate attraverso l'interpolazione quando le osservazioni non sono nè spaziate regolarmente e nè contemporanee, si può dimostrare, Miller (2019), che esse convergono a:

- $\hat{\psi}_M \xrightarrow{d} tr\{(\int W dW' + G^+)'(\int WW')^{-1}(\int W dW' + G^+)(F^+)^{-1}\}$
- $\hat{\rho}_M \xrightarrow{d} (\int Q^2)^{-1}(\int Q dQ + k'G^+k)$
- $\hat{\tau}_M \xrightarrow{d} (k'F^+k \int Q^2)^{-\frac{1}{2}}(\int Q dQ + k'G^+k)$

per $T \to \infty$, dove $\hat{\psi}_M$ è la statistica relativa al test della traccia, $\hat{\rho}_M$ corrisponde a quella del test di Dickey-Fuller basato sui residui e, per concludere, $\hat{\tau}_M$ è quella relativa al t-test. Ci sono diverse quantità che bisogna specificare meglio in queste formule, in quanto molto complesse da ricavare. W è un processo di Wiener, chiamato anche moto Browniano standard, ossia un processo stocastico a tempo continuo che associa ad ogni dato r una variabile casuale reale $W(r)$, tale che $B = \Sigma^{1/2}W$ dove B è un processo simile a W e $\Sigma = \Sigma^{1/2}\Sigma^{1/2'}$ corrisponde alla decomposizione triangolare inferiore di Cholesky. d corrisponde all'ordine di integrazione. Per il calcolo di F^+ e di G^+, invece, dobbiamo supporre che $\Sigma = LL'$ e definire quindi $F^+ = L^{-1'}FL^{-1}$ e $G^+ = L^{-1'}GL^{-1}$, come proposto da Miller (2019). Per quanto riguarda k e Q, invece, seguendo Ghysels e Miller (2014) definiamo $k = (1, k_2')'$ con $k_2 = -(\int W_2W_2')^{-1} \int W_2W_1$ e $Q(r) = W_1(r) - \int W_1W_2'(\int W_2W_2')^{-1}W_2(r) = k'W(r)$ con $W = (W_1, W_2')'$.

3 Modello VAR

3.1 Definizione e rappresentazione generale

Il VAR (Vector Autoregressive Model), modello autoregressivo vettoriale, è uno dei modelli che viene utilizzato maggiormente in ambito statistico sia per l'analisi di serie storiche multivariate, in quanto risulta essere molto flessibile e semplice da applicare, sia per fornire delle previsioni attraverso una teoria basata su delle equazioni che vedremo successivamente. Questa tipologia di modelli è un estensione del modello autoregressivo univariato che viene applicata a serie storiche multivariate e fu introdotta, soprattutto in ambito economico, da Sims nel 1980, il quale dimostrò che questi modelli sono in grado di fornire un insieme di strumenti per poter analizzare delle serie storiche economiche. Questo famoso econometrico americano, come citato da Amisano e Giannini (2012), all'inizio degli anni Settanta si accorse che nei modelli utilizzati fino a quel momento venivano applicati dei metodi che avevano dei punti di debolezza:

- era prevista la specificazione di sistemi di equazioni simultanee basati sull'aggregazione di modelli di equilibrio parziale, senza alcun risultato o riferimento per quanto riguarda le relazioni tra le diverse variabili del sistema (queste invece sono fondamentali quando si analizza la cointegrazione tra variabili);

- la struttura dinamica del modello era spesso specificata in modo da fornire delle restrizioni o regole necessarie per ottenere un'identificazione specifica della forma strutturale dei modelli.

Cerchiamo ora di dare una formula matematica a questa particolare tipologia di modelli e di vedere un esempio per capire come questo va applicato nei casi reali. Sia $Y_t = (y_{1t}, y_{2t}, \ldots, y_{nt})'$ un vettore di n serie storiche. Il modello VAR (p) ha una struttura caratterizzata dal fatto che ogni variabile contenuta in essa è funzione lineare dei ritardi passati della variabile stessa e anche delle altre variabili che vengono analizzate con questo modello. Questa struttura si può rappresentare con l'equazione

$$Y_t = c + \Pi_1 Y_{t-1} + \Pi_2 Y_{t-2} + \cdots + \Pi_p Y_{t-p} + \epsilon_t, \quad t = 1, \ldots, T \qquad (2)$$

dove Π_i (nxn) sono le matrici che contengono i coefficienti e ϵ_t $(nx1)$ è un vettore white noise con media zero e matrice di covarianza Σ.

3.2 Esempio di VAR (1)

Supponiamo ora, ad esempio, di analizzare quattro serie storiche differenti, y_{t-1}, y_{t-2}, y_{t-3}, y_{t-4}. Il modello autoregressivo vettoriale di ordine 1, denominato anche VAR (1), si rappresenta con le seguenti equazioni:

$$y_{t,1} = c_1 + \phi_{11} y_{t-1,1} + \phi_{12} y_{t-1,2} + \phi_{13} y_{t-1,3} + \phi_{14} y_{t-1,4} + \epsilon_{t,1} \tag{3}$$

$$y_{t,2} = c_2 + \phi_{21} y_{t-1,1} + \phi_{22} y_{t-1,2} + \phi_{23} y_{t-1,3} + \phi_{24} y_{t-1,4} + \epsilon_{t,2} \tag{4}$$

$$y_{t,3} = c_3 + \phi_{31} y_{t-1,1} + \phi_{32} y_{t-1,2} + \phi_{33} y_{t-1,3} + \phi_{34} y_{t-1,4} + \epsilon_{t,3} \tag{5}$$

$$y_{t,4} = c_4 + \phi_{41} y_{t-1,1} + \phi_{42} y_{t-1,2} + \phi_{43} y_{t-1,3} + \phi_{44} y_{t-1,4} + \epsilon_{t,4} \tag{6}$$

Si può subito notare come ogni variabile è funzione lineare dei primi ritardi di tutte le variabili nel dataset. Se fosse stato un VAR (2) in ogni singola equazione si sarebbero aggiunti i secondi ritardi di tutte le variabili del dataset. Possiamo giungere alla conclusione che l'ordine del modello VAR, p, ci indica quanti sono i ritardi presenti di ogni singola variabile del nostro sistema da analizzare che vengono utilizzati come predittori per la variabile stessa.

4 Modello CVAR

Il modello CVAR (Cointegrated Vector Autoregressive model), modello autoregressivo vettoriale cointegrato, Kaufmann e Juselius (2013), è utilizzato per analisi statistiche di processi che non sono stazionari. Ad esempio, un processo Random Walk, rappresentato dall'equazione $y_t = y_{t-1} + \epsilon_t$, è una semplice esemplificazione di un processo non stazionario omogeneo di primo grado, riconducibile tramite differenziazione ad un white noise: $\Delta Y_t = Y_t - Y_{t-1} = \epsilon_t$ che risulta essere stazionario. Inoltre, il processo Random Walk può anche essere descritto come il cumulo dei suoi errori fino al tempo t, ossia $x_t = \sum_{i=1}^{t} \epsilon_i + x_0$, dove $\sum_{i=1}^{t} \epsilon_i$ è chiamata trend stocastico. Questo tipo di trend differisce da quello deterministico e fornisce una descrizione più realistica di molte variabili legate al clima, le quali saranno oggetto di studio nell'analisi che faremo successivamente.

I trend stocastici possono essere eliminati sia differenziando la serie sia tramite la cointegrazione (che prevede una combinazione lineare di diverse serie storiche). Tuttavia, la differenziazione rimuove tutte le informazioni a lungo termine dei dati, mentre la cointegrazione assicura che queste siano preservate.

Il modello CVAR cerca di combinare la differenziazione con la cointegrazione per cercare di analizzare, interpretare e descrivere le relazioni a breve termine ma, allo stesso tempo, anche quelle a lungo termine. In questo modo si trova una combinazione che è in grado di risolvere il problema relativo alla differenziazione citato in precedenza.

Cerchiamo ora di dare una formula matematica a questa tipologia di modelli. Kaufmann e Juselius (2013) hanno applicato il CVAR a dati paleoclimatici, suddivisi in:

- variabili climatiche, ad esempio temperatura della superficie terrestre o concentrazione di anidride carbonica nell'aria
- variabili fisiche, ad esempio la concentrazione di ferro
- variabili solari, ad esempio l'insolazione stagionale

Per quanto riguarda il modello, le prime due tipologie di variabili sono considerate endogene, ossia vengono spiegate da altre che appartengono allo stesso modello, mentre l'ultima è considerata esogena, cioè non viene

determinata all'interno del modello bensì ha un valore predeterminato dall'esterno. Detto ciò, l'equazione che lo descrive può essere rappresentata da Kaufmann e Juselius (2013):

$$\Delta x_t = A_0 \Delta w_t + A_1 \Delta w_{t-1} + \Gamma_{11} \Delta x_{t-1} + \Pi(x_{t-1}, w_{t-1}) + \mu_0 + \epsilon_t \qquad (7)$$

dove x_t è un vettore di variabili endogene (quindi climatiche e fisiche), w_t è un vettore di variabili esogene (quindi solari), μ_0 è un vettore di costanti numeriche, A_0, A_1, Γ_{11} e Π sono le matrici che contengono i coefficienti di regressione del modello, Δ rappresenta l'operatore *differenza prima* ossia $\Delta x_t = x_t - x_{t-1}$ e, infine, ϵ_t è un vettore di variabili normali e iid (indipendenti e identicamente distribuite) con media nulla e varianza omega (Ω).

5 Serie storiche irregolari

Lo studio della realtà economica e finanziaria prevede l'utilizzo di vari strumenti statistici per descrivere e interpretare la dinamica temporale di vari fenomeni disponibili sotto forma di serie storica. Alcuni esempi possono essere le serie mensili riguardanti le vendite di una determinata azienda, i dati trimestrali sul numero di clienti di un certo negozio, le vendite annuali di una concessionaria e tantissime altre tipologie.

5.1 Descrizione

Le analisi di serie storiche multivariate, solitamente, prevedono lo studio e l'interpretazione di dati che vengono osservati regolarmente e alla stessa frequenza. Queste particolari caratteristiche non sono presenti nella nostra applicazione, in quanto le serie storiche analizzate non sono osservate nè regolarmente, nè contemporaneamente, e sono un esempio di serie storiche a frequenza mista, ossia caratterizzate dal fatto che sono osservate a frequenze differenti (perciò non tutte le osservazioni sono osservate contemporaneamente). Inoltre, esse sono fisiche, ossia traggono origine dall'osservazione e dallo studio di fenomeni metereologici, e sono formate da più variabili, motivo per cui vengono chiamate serie storiche multivariate o multiple.

L'analisi che descriverò prevede lo studio di due o più serie storiche spaziate irregolarmente e con poche, o addirittura quasi zero, osservazioni contemporanee, e questo è dovuto alla difficoltà di reperire dati paleoclimatici, quali ad esempio le concentrazioni di anidride carbonica (CO_2) nell'aria e le temperature della superficie terrestre.

Come citato da Miller (2019), ci sono diversi casi che si possono analizzare, tra cui:

1. due o più serie che vengono osservate a frequenze regolari ma differenti
2. una singola serie che cambia la frequenza delle osservazioni durante il periodo di osservazione

Miller (2019) introduce delle metodologie che possono essere applicate per analizzare la cointegrazione tra serie storiche irregolari o non contemporanee. Supponiamo di avere due serie storiche multivariate, α_t^* e β_t^*, osservate agli istanti X_p e Y_q rispettivamente, e che siano osservate solamente α_{X_p} e β_{Y_q},

dove $\alpha_{X_p} = \alpha^*_{X_p}$ e $\beta_{Y_q} = \beta^*_{Y_q}$. Con queste ipotesi non ci assicuriamo che X_p sia uguale a Y_q per ogni q o per ogni p, e quindi è difficile applicare i test di cointegrazione su delle serie storiche con queste caratteristiche. Perciò i dati devono essere modificati in qualche modo.

5.2 Metodi per le analisi

Ci sono varie strategie che possono essere utilizzate per affrontare il problema dei dati irregolari o spaziati irregolarmente. Ryan e Giles (1999) ne considerano 3 che sono:

- ignorare completamente le osservazioni mancanti;
- rimpiazzare le osservazioni mancanti con l'ultima disponibile;
- utilizzare il metodo dell'interpolazione lineare per sostituire le osservazioni che mancano.

Tuttavia queste tecniche non sono sempre adatte al nostro caso, in quanto falliscono se X_p è diverso da Y_q, ossia se le due serie osservate non corrispondono.

Due dei metodi più comuni che vengono utilizzati per analizzare queste tipologie di serie storiche sono l'interpolazione lineare e l'interpolazione *a step*. Miller (2019) definisce una serie storica interpolata *a step* ($\hat{\alpha}^S_t$) come $\hat{\alpha}^S_t = \alpha_{X_p}$ per $t \in [X_p, X_{p+1}]$, mentre una serie storica interpolata linearmente ($\hat{\alpha}^L_t$) come $\hat{\alpha}^L_t = \alpha_{X_p} + \frac{t-X_p}{X_{p+1}-Xp}(\alpha_{X_{p+1}} - \alpha_{X_p})$ per $t \in [X_p, X_{p+1}]$. Come possiamo notare, la seconda equazione (relativa all'interpolazione lineare) contiene la prima equazione (relativa all'interpolazione *a step*). Ciò che le distingue è che al secondo membro della seconda equazione è sommata una quantità che è il risultato del prodotto tra una frazione, ottenuta come rapporto tra due quantità ricavate sottraendo da entrambe X_p, e la differenza tra le serie ai tempi che sono esattamente gli estremi dell'intervallo considerato.

Per fissare le idee sui due metodi, Miller (2019) suppone di avere dei dati non contemporanei ma spaziati regolarmente in n unità, in modo che $q = p$ e $m = X_{p+1} - X_p = Y_{p+1} - Y_p$, e considera X_p il giovedi della settimana p mentre Y_q il sabato della stessa settimana. In questo caso la serie α_{X_p} è

interpolata come

$$\hat{\alpha}_t^S = \hat{\alpha}_{X_{p+i}}^S = \alpha_{X_p} \qquad \text{per i} = 0, \ldots, \text{m - 1} \quad \text{oppure} \tag{8}$$

$$\hat{\alpha}_t^L = \hat{\alpha}_{X_{p+i}}^L = \alpha_{X_p} + \frac{i}{m}(\alpha_{X_{p+1}} - \alpha_{X_p}) \qquad \text{per i} = 0, \ldots, \text{m} \tag{9}$$

usando rispettivamente l'interpolazione *a step* e quella lineare. Di ciascuna delle due serie interpolate possiamo rappresentare l'equazione delle differenze prime, ossia

$$\Delta\hat{\alpha}_{X_{p+i}}^S = \begin{cases} \alpha_{X_p} - \alpha_{X_{p-1}} & \text{per i} = 0 \\ 0 \qquad \text{per i} = 1, \ldots, \text{m-1} \end{cases} \tag{10}$$

$$\Delta\hat{\alpha}_{X_{p+i}}^L = \frac{1}{m}(\alpha_{X_{p+1}} - \alpha_{X_p}) \qquad \text{per i} = 1, \ldots, \text{m} \tag{11}$$

Il sistema relativo all'interpolazione *a step* mostra come la differenza prima assuma sempre il valore zero, tranne nel caso in cui i sia uguale a 0 dove essa assume come valore la differenza tra la serie osservata al tempo p e la serie osservata al tempo $p-1$. Per quanto riguarda l'interpolazione lineare, invece, la differenza prima assume sempre lo stesso valore a prescindere da quanto vale l'indice i. Questo valore è il risultato del prodotto tra una costante, $\frac{1}{m}$, e la differenza tra il valore assunto dalla serie al tempo $p+1$ e quello al tempo p. Si può anche notare come nel primo caso si tenga conto dell'osservazione al tempo p e quella precedente, mentre nel secondo caso si dà maggiormente importanza all'osservazione futura rispetto al tempo p, ossia al tempo $p+1$.

6 Applicazione ai dati paleoclimatici

L'analisi di serie storiche di tipo paleoclimatico è fondamentale nella nostra epoca in quanto è sempre più evidente, soprattutto dagli eventi atmosferici che si stanno verificando di recente, che il nostro pianeta sta attraversando un periodo dove il fenomeno del *riscaldamento globale* è sempre più forte. Esso causa lo scioglimento dei ghiacciai, terremoti, frane continue di catene montuose e altri danni ambientali, perciò risulta di fondamentale importanza cercare di studiare, analizzare ed interpretare dei dati relativi al clima in generale. Uno dei principali problemi che si riscontra quando si inizia a cercare questi dati è il metodo di raccolta. Infatti, risalire a dati di tipo paleoclimatico non è un processo facile, e i dati che ne derivano sono, proprio a causa di questa difficoltà di reperimento, irregolari e difficili da trattare, con molte osservazioni mancanti e quindi le serie storiche che ne derivano non possono essere trattate come quelle regolari, le quali possono essere analizzate abbastanza facilmente.

Kaufmann e Juselius (2013) hanno analizzato un sistema formato da alcune serie storiche di dati paleoclimatici e, successivamente, stimato un CVAR per capire quali sono i meccanismi che guidano i cicli glaciali durante l'epoca precedente a quella attuale (Olocene), chiamata Pleistocene o periodo di Vostok. Nella scala dei tempi geologici quest'ultima è la prima delle due epoche che costituisce il periodo quaternario, ed è compresa tra 2,58 milioni di anni fa e 11700 anni fa.

Ci sono molte serie storiche irregolari che vengono analizzate, tra cui la quantità di diossido di carbonio presente nell'aria, la quantità di metano (CH_4) presente nell'aria e il volume dei ghiacciai. Kaufmann e Juselius (2013) hanno scoperto che esiste cointegrazione tra la serie dell'anidride carbonica presente nell'aria e quella della temperatura della superficie terrestre. Secondo la comunità scientifica l'anidride carbonica è una delle principali responsabili del cambiamento climatico degli ultimi anni, perciò è ragionevole pensare che esista cointegrazione tra le due serie citate precedentemente.

Come citato nel paragrafo precedente, le serie che sono risultate essere cointegrate tra di loro sono state quelle della temperatura e quella della concentrazione di anidride carbonica. In figura 1 e 3 possiamo notare, rispettivamente, la serie della concentrazione di CO_2 di Luthi et al. (2008) e quella della temperatura di Jouzel et al. (2007), le quali sono state inserite

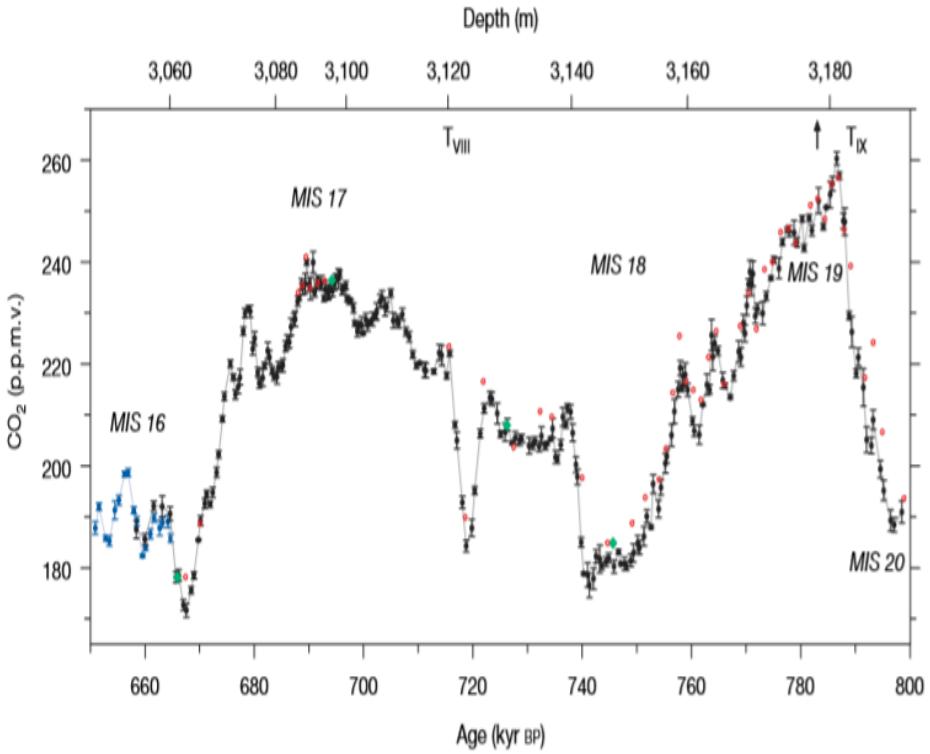

Figura 1: Concentrazione di CO_2 al variare della profondità e del tempo, Luthi et al. (2008)

nell'articolo per mostrare come sono costituite le serie oggetto di studio della nostra analisi, ma molto difficili da interpretare e da spiegare, in quanto sono costruite seguendo una scala temporale chiamata EDC3 che, come citato da Parrenin et al. (2007), è basata su dei fenomeni atmosferici di diversa entità, ad esempio cumuli di neve, ma anche le eruzioni vulcaniche avvenute durante l'ultimo millennio, i quali sono molto complessi da analizzare. Tuttavia, per effettuare queste misurazioni viene utilizzata questa scala temporale, in quanto essa ci garantisce una grande accuratezza sulla durata degli eventi, con una misura di incertezza pari a circa 6000 anni su 800000, che corrisponde ad un 0,75 % di errore e che sembra essere una misura di incertezza davvero ottima per l'analisi di cui ci occupiamo, dato che i dati che possediamo arrivano fino a circa 800000 anni fa. Le rilevazioni sono state effettuate in una zona dell'Antartide, chiamata Dome C.

La figura 1 contiene il grafico relativo all'andamento della concentrazione

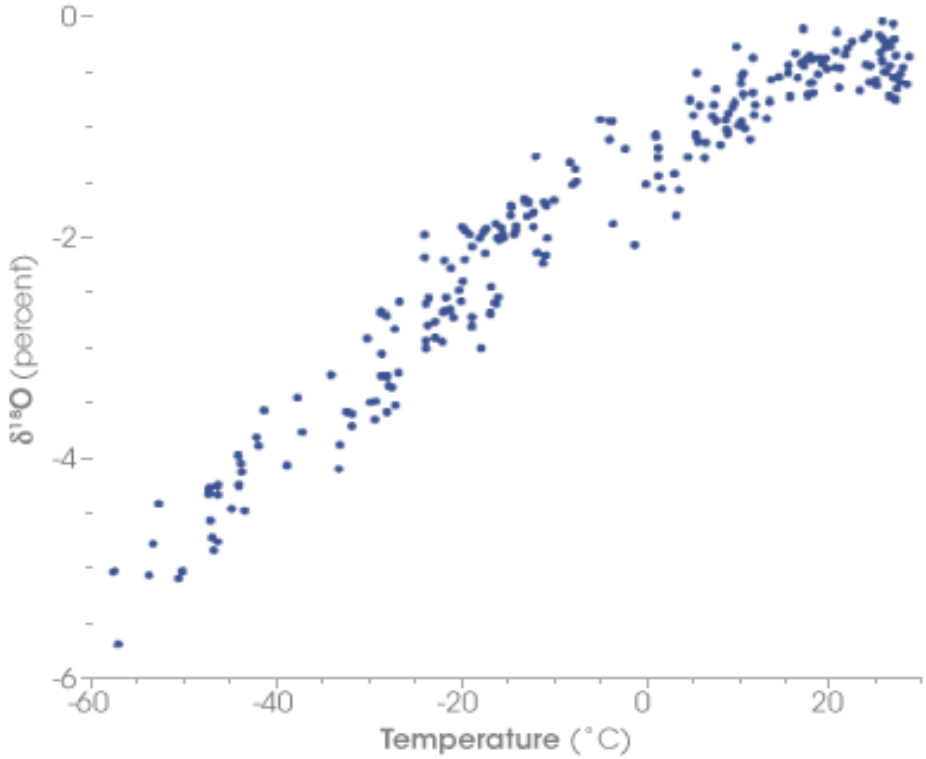

Figura 2: Andamento di *Delta-O-18* al variare della temperatura, Riebeek (2005)

di anidride carbonica, misurata in ppmv (parts per million by volume), al variare della profondità (*Depth*, misurata in metri) che varia da circa 3040 a 3200 metri, e del tempo (*Age*, misurata in 'kyr', ossia 'migliaia di anni fa rispetto al presente') che varia da 650000 fino a 800000 anni fa. Come si può subito notare, il grafico contiene delle piccole circonferenze di colore nero collegate da una linea grigia e questi sono i dati misurati dall'Università di Berna (Svizzera), i quali contengono anche delle barrette che corrispondono alla deviazione standard dalla media. Tuttavia, il grafico non contiene solo la serie storica corrispondente all'andamento della CO_2 ma anche alcuni simboli, motivo per cui nel paragrafo precedente si dice che le serie storiche oggetto di studio hanno una certa difficoltà ad essere interpretate e capite. Questi simboli sono una misura di incertezza delle osservazioni ottenute, ad esempio i cerchietti rossi rappresentano i dati misurati in un laboratorio di geofisica dell'Università di Grenoble (Francia) con un'incertezza di 2σ

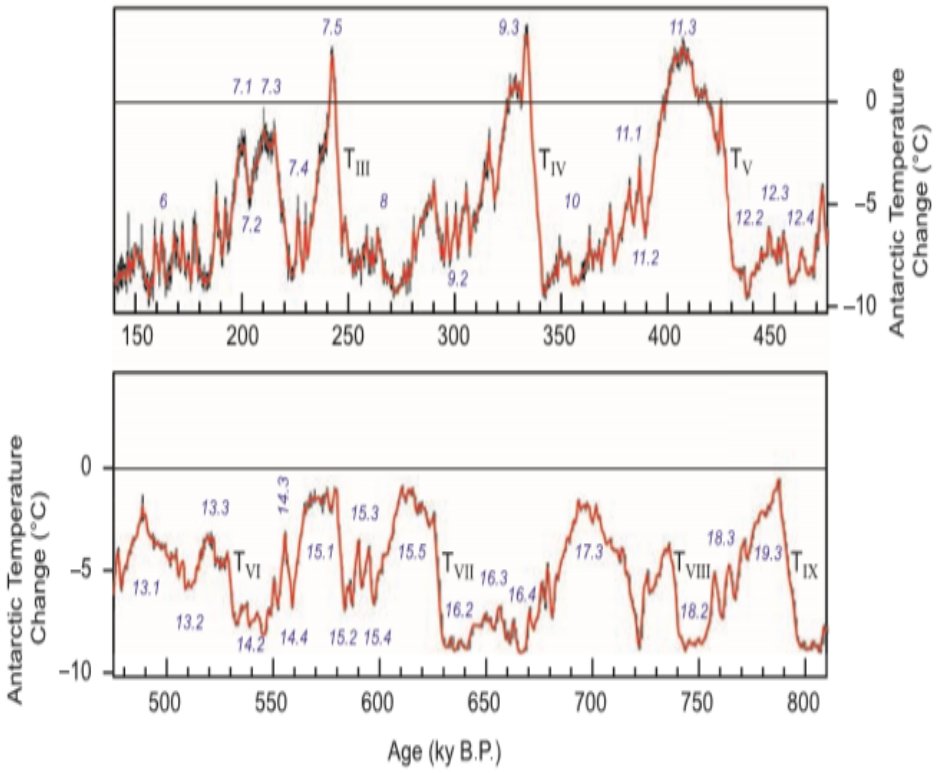

Figura 3: Andamento della temperatura nel tempo fino a 800000 anni fa, Jouzel et al. (2007)

che corrisponde a 3 ppmv. Gli altri simboli sono tutte misurazioni ricavate attraverso delle tecniche molto complesse di estrazione dell'aria.

La figura 3, invece, mostra l'andamento nel tempo della temperatura antartica a partire da circa 810000 anni fa fino ad arrivare a 140000 anni fa (i due grafici sono spezzati in due immagini differenti solo per poter osservare meglio i dati contenuti). La prima cosa che si nota è che, sull'asse delle ordinate, i valori vanno da zero a -10. Questi valori stanno ad indicare in che termini è cambiata la temperatura della zona antartica negli anni passati. Si può vedere, ancora una volta, come la temperatura sia rappresentata da un andamento altalenante, con variazioni positive e negative più o meno significative. I valori che sono riportati corrispondono alla media calcolata ogni 100 anni oppure alla media di periodi più corti, come citato da Jouzel et al. (2007).

I dati relativi alla temperatura e al diossido di carbonio vengono ricavati

dalle carote di ghiaccio, le quali si possono definire come delle sezioni di ghiaccio che si ottengono attraverso una tecnica, chiamata carotaggio, che prevede la perforazione dei ghiacciai a delle profondità regolari. Nonostante si possa pensare che delle sezioni di ghiaccio non siano così utili da analizzare, esse possono fornire delle indicazioni su diversi parametri atmosferici, tra i quali le radiazioni solari, la composizione dell'aria (quindi anche la presenza di anidride carbonica) e, addirittura, eventi straordinari come terremoti o eruzioni vulcaniche.

Per quanto riguarda le misurazioni relative alla temperatura, risultano di fondamentale importanza l'*Oxygen-18* e il *delta-O-18*. Riebeek (2005) descrive il primo come *"Un isotopo di ossigeno che ha due neutroni extra, per un totale di 10 neutroni e 8 protoni, a differenza di un atomo normale di ossigeno che possiede 8 neutroni e 8 protoni"*. Questa tipologia di atomi è molto importante perchè permette agli scienziati di analizzare le concentrazioni di ossigeno presenti nell'aria e, quindi, di poter studiare i cambiamenti climatici del passato. Il secondo ($\delta^{18}O$), invece, è una misura utilizzata per analizzare le temperature di precipitazione. La figura 2, citata da Riebeek (2005), contiene un grafico che rappresenta l'andamento della percentuale di *Delta-O-18* presente nelle precipitazioni al variare della temperatura rispetto a quella che ci sarebbe nel caso si registrasse la temperatura media annuale. Si può immediatamente notare come il grafico abbia un andamento lineare e che la percentuale di *Delta-O-18* aumenti all'aumentare della temperatura. Possiamo anche vedere che a temperature estremamente basse, corrispondenti alle zone come Siberia, Groenlandia ed altre, la percentuale presente sia del 5 percento in meno rispetto all'acqua degli oceani; dall'altro lato, invece, si vede come a temperature estremamente elevate, corrispondenti alle zone come l'Australia, la percentuale presente sia praticamente pari a quella delle acque degli oceani.

Le temperature che vengono utilizzate nella nostra analisi sono espresse in gradi *Celsius* e come differenze dalla temperatura media verificatasi negli ultimi 100000 anni (nel nostro caso si fa rimerimento al 'presente' come l'anno 1950, quindi si intende gli ultimi anni fino al 1950). Nell'analisi che segue, come citato da Miller (2019), sono disponibili 5788 osservazioni tra 801662 anni fa e 38 anni fa.

Per quanto riguarda le concentrazioni di anidride carbonica, invece, esse sono espresse con l'unita di misura *ppmv*, acronimo di *parts per million by vo-*

lume, la quale viene solitamente utilizzata per esprimere errori di misurazione ma, soprattutto, per esprimere livelli estremamente bassi di concentrazione di un elemento chimico, come nel nostro caso. Nell'analisi che descrivo, come citato da Miller (2019), ci sono 1096 osservazioni disponibili tra 798512 anni fa e 137 anni fa. Inoltre, le concentrazioni di CO_2 lungo tutta la serie variano da un valore minimo di 171,6 ppmv fino ad un valore massimo di 298,6 ppmv, perciò hanno una variazione massima di 127 ppmv. Questo ci conferma il fatto che l'andamento nel tempo delle concentrazioni di CO_2 non sia un fenomeno regolare e non segua un determinato trend. Un altro fenomeno da tenere in considerazione è il fatto che questo particolare tipo di dato paleoclimatico ha avuto un aumento da circa 285 ppmv (verso l'anno 1850) a 406 ppmv (nel 2017), con un incremento di 121 ppmv in circa 170 anni.

Kaufmann e Juselius (2013) stimano un modello CVAR applicato a varie serie storiche paleoclimatiche per cercare di capire quali sono i meccanismi che guidano i cicli glaciali durante l'era geologica chiamata Pleistocene. Per analizzare la cointegrazione tra queste serie storiche viene utilizzato il metodo dell'interpolazione lineare (approfondito nella sezione relativa ai metodi per l'analisi di serie storiche irregolari), in quanto esso risulta essere adatto per trattare serie costituite da dati irregolari.

6.1 Simulazioni di Monte Carlo

Per approfondire l'analisi sui dati paleoclimatici Miller (2019) utilizza un esperimento di Monte Carlo, per cercare di esaminare e valutare da un punto di vista teorico i modelli applicati a serie storiche irregolari.

Il primo passo di Miller (2019) consiste nel simulare la serie α_t^* facendo in modo che la varianza dei suoi incrementi corrisponda alla varianza campionaria della quantità $(\alpha_{X_p} - \alpha_{X_{p-1}})/\sqrt{X_p - X_{p-1}}$ per quanto riguarda il dato paleoclimatico relativo alla concentrazione di anidride carbonica nell'aria. Come si può notare, questa quantità è il risultato del rapporto tra la differenza delle serie al tempo p e $p-1$ (quindi come fosse una differenza prima) e la radice quadrata della differenza dei tempi in cui queste serie sono osservate.

Successivamente l'Autore simula la seconda serie storica, ossia β_t^*, usando il modello

$$\beta_t^* = \hat{\gamma} + \hat{\delta}\hat{\alpha}_t^* + \epsilon_t^* \tag{12}$$

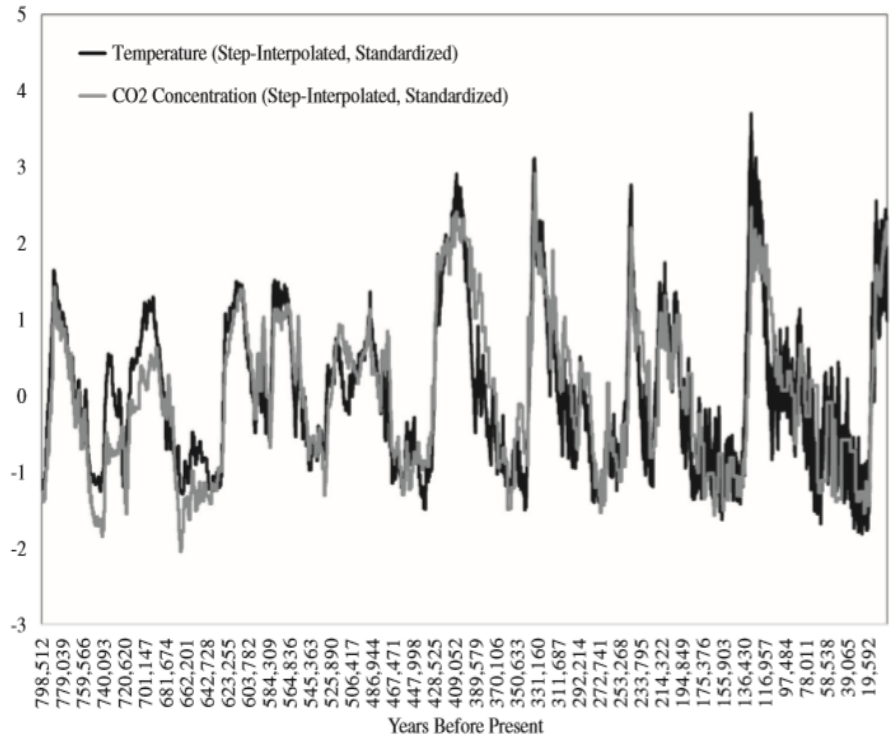

Figura 4: Temperatura e CO_2 con interpolazione *a step*, Miller (2019)

con $\hat{\gamma} = -13,6346$ e $\hat{\delta} = 0,0490$, valori che vengono stimati utilizzando l'interpolazione *a step*. Possiamo notare che il modello in equazione (12) sembra avere la stessa struttura di uno di regressione lineare con $\hat{\gamma}$ costante di regressione, β_t^* variabile dipendente oggetto di studio, $\hat{\delta}$ coefficiente di regressione, $\hat{\alpha}_t^*$ variabile dipendente e ϵ_t^* termine d'errore.

Sotto il primo processo generatore dei dati (DGP #1), come citato da Miller (2019), la concentrazione di anidride carbonica e la temperatura non risultano essere cointegrate. Di conseguenza, il termine del modello ϵ_t^* è generato come un random walk (vedi sezione 4 per la rappresentazione) con una varianza di 0,0035. Questo preciso valore è stato scelto in modo che, come nel caso della prima serie simulata, la varianza degli incrementi di β_t^* corrisponda alla varianza campionaria della quantità $(\beta_{Y_q} - \beta_{Y_{q-1}})/\sqrt{Y_q - Y_{q-1}}$.

Sotto il secondo processo generatore dei dati (DGP #2), invece, Miller (2019) considera le due serie cointegrate. Di conseguenza, il termine del modello $\epsilon_t^* = (1 - \rho_T^2)\mu_t + \sigma_v \nu_t$, dove $\mu_t = \rho_T \mu_{t-1} + w_t$ con $\rho_T = (1 -$

$15/T) = 0,999981$, $\sigma_v^2 = 0,001761$ e il vettore (w_t, ν_t) normale standard. Il valore relativo al parametro σ_v^2 è stato scelto per lo stesso motivo utilizzato nel primo processo generatore dei dati, mentre il valore del parametro ρ_T è stato scelto per corrispondere a una stima del parametro di un modello autoregressivo stimato utilizzando l'interpolazione *a step*.

Le serie storiche annuali simulate α_t^* e β_t^* successivamente subiscono una modifica, ossia non vengono analizzate le intere serie ottenute, bensì si tengono in considerazione soltanto le osservazioni che corrispondono a quelle che vengono osservate, rispettivamente, per i dati relativi ai due fenomeni oggetto di studio, ossia la concentrazione di anidride carbonica nell'aria e la temperatura. In questo modo si ottengono delle simulazioni che sono utili quando vengono analizzate delle serie storiche irregolari, non contemporanee, spaziate irregolarmente oppure altre tipologie simili.

Dopo aver simulato le due serie, Miller (2019) a quelle nuove ottenute dopo la modifica applica i metodi di:

- interpolazione lineare
- interpolazione *a step*
- standardizzazione

e successivamente i test su di esse vengono effettuati sulla serie originaria e su quella ritardata di un lag per controllare se è presente correlazione seriale tra le serie oggetto di analisi. Come citato da Miller (2019), se vengono osservati i dati annuali non è necessario applicare alcuna differenza prima in quanto possiamo utilizzare l'osservazione relativa a questa tipologia di dato.

La figura 4 rappresenta le serie standardizzate relative alla temperatura e alla concentrazione di anidride carbonica nell'aria. Il grafico contiene sull'asse delle ascisse il tempo, rappresentato in anni trascorsi dal presente (che nel nostro caso corrisponde all'anno 1950) che vanno da 19592 anni fa fino a 798512 anni fa, mentre sull'asse delle ordinate le unità standardizzate e analizzate attraverso l'interpolazione *a step*. Questo grafico non ha un andamento particolare (come ad esempio quello lineare o parabolico), ma molto altalenante, con variazioni positive che sono più significative rispetto a quelle negative. Infatti per quanto riguarda quelle positive ci sono unità che arrivano fino ad un valore superiore a 2 (rispetto alla media), mentre quelle negative arrivano fino ad un valore di 1,5 rispetto alla media (sembra ci sia

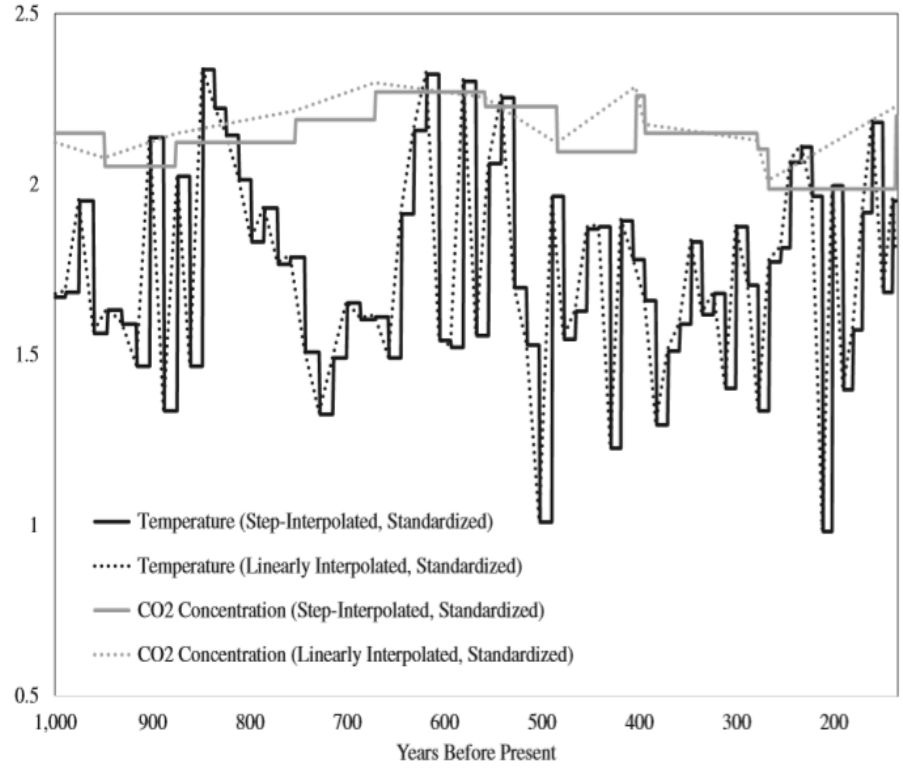

Figura 5: Temperatura e CO_2 con interpolazione *a step* e lineare, Miller (2019)

solo una singola osservazione che superi il valore 2). Questo dimostra come sia più probabile che la temperatura sia superiore alla media. Le due serie risultano essere quasi sovrapposte per la prima parte corrispondente alle osservazioni meno recenti, per poi avere un andamento abbastanza simile nel resto del grafico, ad esclusione della parte finale corrispondente alle osservazioni più recenti, dove le due serie sembrano assumere valori abbastanza diversi.

La figura 5, invece, rappresenta le serie standardizzate relative ai due fenomeni oggetto di studio ottenute attraverso la standardizzazione e analizzate con i metodi dell'interpolazione lineare e interpolazione *a step*. Il grafico contiene sull'asse delle ordinate le unità standardizzate come in figura 4, mentre su quello delle ascisse il tempo che varia da 1000 anni fa fino a circa 150 anni fa. Si può subito notare come le serie relative alla temperatura analizzata attraverso le due tipologie di interpolazione siano nettamente dif-

ferenti rispetto a quelle relative alla concentrazione di anidride carbonica, in quanto la prima è per la maggior parte ad un livello più basso rispetto alla seconda e quest'ultima invece, rispetto alla prima che ha un andamento molto altalenante con valori che vanno da circa 1 a 2,2, risulta essere molto stabile nel tempo, con dei valori che variano da circa 1,9 a 2,3. Questo lo possiamo interpretare con il fatto che, negli ultimi anni fino al 1950, la concentrazione di anidride carbonica subiva piccole variazioni, mentre la temperatura poteva variare anche quasi di un punto (vedi osservazione relativa circa a 200 anni fa). Se volessimo prevedere un grafico relativo all'andamento di queste due tipologie di serie relative agli anni più recenti, possiamo pensare che sia ragionevole avere dei valori maggiori di temperatura, dovuti principalmente al fenomeno del 'riscaldamento globale', mentre per quanto riguarda l'anidride carbonica potrebbe esserci una leggera diminuzione in futuro, in quanto vengono introdotte nuove tecnologie che cercano di diminuire l'inquinamento globale, come ad esempio le auto ibride o elettriche.

Dopo aver analizzato e commentato le figure contenenti i grafici relativi ai vari metodi di analisi di serie storiche irregolari applicati alle serie paleoclimatiche considerate, ci occupiamo ora delle simulazioni di Monte Carlo e vediamo i test statistici a che conclusioni ci portano per quanto riguarda la cointegrazione. La tabella 1, come citato da Miller (2019), contiene le probabilità di rifiuto relative a 1000 simulazioni di Monte Carlo. $\hat{\psi}_T(j)$ è relativa al test della traccia con ipotesi nulla "presenza di j trend stocastici" (che è equivalente a dire 2 - j relazioni di cointegrazione) e con ipotesi alternativa "presenza di una quantità di trend stocastici minore a j". Per quanto riguarda la colonna p, se essa è uguale a zero significa che è inclusa la serie storica originaria (quindi con zero ritardi), mentre se è uguale a uno è inclusa la serie storica ritardata di un lag. Le etichette relative alle righe, invece, rappresentano i risultati ottenuti con le due tipologie di interpolazione che consideriamo nella nostra analisi, ossia quella lineare (L) e quella *a step* (S). La riga chiamata "CV", infine, contiene dei valori corrispondenti a un valore nominale di 0,05 (quello utilizzato più spesso in ambito statistico) che vengono ricavati da Johansen e Juselius (1990) e Phillips e Ouliaris (1990).

I risultati sono descritti in tabella 1. I valori contenuti nelle colonne denominate con $\hat{\psi}_T(1)$ e $\hat{\psi}_T(2)$ si differenziano per il fatto che:

- $\hat{\psi}_T(1)$ riguarda il test ha ha come ipotesi nulla la presenza di cointegrazione, contro l'alternativa di cointegrazione non presente

Tabella 1: Probabilità di rifiuto con 1000 simulazioni di Monte Carlo, Miller (2019)

DGP #1	p	$\hat{\psi}_T(1)$	$\hat{\psi}_T(2)$	$\hat{\rho}_T$	$\hat{\tau}_T$
$\hat{z}_t^L$	0	0,658	0,995	0,000	0,247
	1	0,004	0,021	0,949	0,018
$\hat{z}_t^S$	0	0,004	0,047	0,068	0,063
	1	0,004	0,047	0,069	0,064
DGP #2	p	$\hat{\psi}_T(1)$	$\hat{\psi}_T(2)$	$\hat{\rho}_T$	$\hat{\tau}_T$
$\hat{z}_t^L$	0	0,212	0,979	0,000	0,000
	1	0,035	0,967	1,000	0,998
$\hat{z}_t^S$	0	0,029	1,000	1,000	1,000
	1	0,029	1,000	1,000	1,000

- $\hat{\psi}_T(2)$ riguarda il test con ipotesi nulla corrispondente alla non cointegrazione, contro l'alternativa di possibile cointegrazione

Come citato da Miller (2019), nel caso del primo processo generatore dei dati (DGP #1) consideriamo i valori relativi a $\hat{\psi}_T(2)$, $\hat{\rho}_T$ e $\hat{\tau}_T$ come la dimensione empirica dei test, mentre con il secondo processo generatore dei dati (DGP #2) consideriamo il valore corrispondente a $\hat{\psi}_T(1)$. Nel caso del secondo processo generatore dei dati, invece, consideriamo le colonne relative a $\hat{\psi}_T(2)$, $\hat{\rho}_T$ e $\hat{\tau}_T$ come la potenza dei test. Prima di andare ad approfondire ogni tipo di considerazione sui risultati della tabella chiariamo brevemente cosa intendiamo per *dimensione empirica del test* e *potenza del test*.

Quando prendiamo in considerazione la dimensione empirica di un test statistico ci riferiamo a una misura di robustezza di quel test. Supponiamo ad esempio di applicare una soglia di rifiuto del 5 %. Ciò che mi aspetto è che, eseguendo questo test su molti campioni legati dalla stessa distribuzione di riferimento sotto l'ipotesi nulla, il 5 % di questi campioni mi porteranno a rifiutare l'ipotesi nulla quando questa è vera, ossia a commettere un errore di primo tipo, in quanto sono stati generati proprio da questa ipotesi. Per vedere come varia la dimensione empirica del test si possono far variare alcune quantità, come la numerosità del campione n, il numero di ritardi considerati p o anche altre, e si ripete il test per molti campioni. Se la dimensione empirica del test rimane invariata al variare di queste quantità allora esso risulta essere robusto, altrimenti c'è qualcosa che non funziona.

Per quanto riguarda la potenza del test, invece, essa può essere definita come la probabilità di rifiutare correttamente l'ipotesi nulla e può assumere valori compresi tra zero e uno, dove più il valore si avvicina a 1 e più significa che il nostro test è potente, ossia fa la scelta giusta di fronte alla possibilità di rifiutare o non rifiutare l'ipotesi nulla. Ovviamente, dall'altro lato, più la potenza si avvicina a 0 e più il nostro test diventa distorto, ossia prende la scelta sbagliata nella maggior parte dei casi.

Dopo aver spiegato brevemente i due dati che andremo a considerare nei test passiamo ora ad occuparci di quali conclusioni ci porta ogni singolo valore che vediamo in tabella 1. Se confrontiamo i valori contenuti nella tabella con il valore nominale di 0,05, ossia quello che si tiene solitamente come riferimento per decidere se rifiutare o non rifiutare l'ipotesi nulla, usando il metodo dell'interpolazione lineare e la serie originaria con zero ritardi si può notare che la dimensione empirica dei test (citata nel paragrafo precedente) è piuttosto distorta. Infatti, essa vale 0,995, 0,000, 0,247, 0,212 e questi valori risultano molto diversi dal valore nominale di 0,05, e perciò il test effettuato in questo caso non è propriamente dei migliori. Tuttavia, se considerassimo la potenza del test vediamo che $\hat{\psi}_T(2)$ è molto vicina a 1 (0,979) e ciò significherebbe aver preso la scelta giusta nella maggior parte dei casi, ma $\hat{\rho}_T$ e $\hat{\tau}_T$ risultano pari a zero, e quindi questi ci portano a considerare il test non robusto. Se consideriamo l'interpolazione *a step* applicata alla serie originaria, invece, otteniamo dei valori pari a 0,047, 0,068, 0,063 e 0,029. Rispetto al caso precedente, il test effettuato risulta essere robusto in quanto i valori ottenuti sono molto vicini a quello nominale di 0,05. A conferma di questo possiamo osservare che $\hat{\psi}_T(2)$, $\hat{\rho}_T$ e $\hat{\tau}_T$, corrispondenti alla potenza del test, sono tutte pari a 1 e questa è la condizione ottimale per capire se il test è robusto oppure no.

Consideriamo ora il caso in cui la serie sia ritardata di un lag. Nel caso dell'interpolazione lineare possiamo notare subito come, rispetto all'analisi della serie con p = 0, i risultati ottenuti siano notevolmente migliori. Infatti, i 4 valori relativi alla dimensione empirica risultano essere 0,021, 0,949, 0,018 e 0,035. Solo 0,949 è un valore non accettabile in quanto molto diverso da quello nominale di 0,05, mentre gli altri tre sembrano essere tutto sommato buoni come risultati. A conferma di ciò, possiamo guardare la potenza del test e vediamo che $\hat{\psi}_T(2)$, $\hat{\rho}_T$ e $\hat{\tau}_T$ assumono rispettivamente i valori 0,967, 1,000 e 0,998 che sono tutti praticamente pari (o quasi) a 1 e, rispetto al caso

analizzato nel paragrafo precedente, possiamo dire che ritardare la serie di un lag porta notevoli miglioramenti per quanto riguarda il test della traccia effettuato. Nel caso dell'interpolazione *a step*, invece, non ci sono grandissimi cambiamenti come in quella lineare. Infatti, applicare il metodo alla serie ritardata di un lag porta a ottenere dimensioni empiriche pari a 0,047, 0,069, 0,064 e 0,029 e questi valori corrispondono esattamente a quelli ottenuti con la serie senza ritardi. Come nel caso precedente, tutti e tre i valori relativi alla potenza del test sono pari a 1, a conferma che anche in questo caso il test è robusto.

6.2 Risultati empirici

Nonostante sia noto che il metodo dell'interpolazione causi molto spesso dei problemi con la potenza e la dimensione empirica dei test statistici, nel nostro caso conduce a dei risultati molto buoni rispetto a ciò che ci si potrebbe aspettare. Infatti, le simulazioni effettuate con i dati paleoclimatici dimostrano come, sia nel caso con $p = 0$ che in quello con $p = 1$, i test risultano essere robusti e avere una potenza davvero ottima. Vediamo ora nello specifico le diverse analisi effettuate e le conclusioni che ci danno i risultati.

Dai risultati ottenuti in precedenza attraverso il metodo di Monte Carlo, si può dire che applicare l'interpolazione *a step* per analizzare questa tipologia di serie storiche è sicuramente più vantaggioso rispetto a quella lineare. Infatti, nel primo caso i test risultano essere robusti quando viene applicata sia la serie originaria che quella ritardata di un lag, mentre nel secondo i valori che otteniamo risultano essere distorti sia nel caso dell'applicazione alla serie originaria, sia in quello alla serie ritardata di un lag. Bisogna comunque riconoscere il fatto che ritardare la serie di un lag nel secondo caso porta dei notevoli miglioramenti, anche se non tali da considerare il test robusto.

Le conclusioni che abbiamo tratto dalla precedente analisi, relative al fatto che l'interpolazione lineare non è adatta a trattare serie storiche di tipo irregolare, sono sostenute anche da:

- Ryan e Giles (1999) che hanno applicato dei test per verificare la presenza di radici unitarie in serie storiche con osservazioni mancanti e hanno utilizzato diverse strategie per provare a trattarle, tra cui il metodo dell'interpolazione lineare che si è rivelato non adatto a questa tipologia di test

- Ghysels e Miller (2014) che hanno applicato dei test di cointegrazione a dei dati analizzati attraverso l'interpolazione lineare e scoprono che con questo metodo c'è un'evidente distorsione nei risultati ottenuti

In tabella 2 sono presenti dei valori relativi a dei test applicati a un sistema paleoclimatico bivariato. Con $\hat{\psi}_j(j)$ è rappresentato un test della traccia con ipotesi nulla che corrisponde alla presenza di j trend stocastici (quindi 2 - j relazioni di cointegrazione) contro l'alternativa corrispondente alla presenza di una quantità di trend stocastici minore di j. p rappresenta il numero di ritardi con cui è ritardata la serie (quindi ad esempio p = 0 è la serie originaria, p = 10 è la serie ritardata di dieci lag). Le etichette relative alle righe, come in tabella 1, rappresentano rispettivamente i risultati ottenuti attraverso i metodi dell'interpolazione lineare (L) e quella *a step* (S). La riga chiamata "CV" corrisponde esattamente a quella che si trova anche in tabella 1, arrotondata questa volta alla seconda cifra decimale.

Questa tabella mostra ciò che si ottiene includendo dalla serie originaria fino alla serie ritardata di 50 lag, con un salto di 1 lag per ciascuna serie. Il metodo appena citato viene utilizzato per verificare effettivamente la robustezza del test che si applica alle serie storiche analizzate. Inoltre, la scelta di tenere 50 lag è dovuta anche al fatto che le conclusioni che si otterrebbero con le serie ritardate di un lag (p = 1) utilizzando l'interpolazione lineare e quella *a step* sarebbero in evidente contrasto tra di loro e renderebbero difficile la nostra analisi, perciò utilizzando 50 lag si evitano eventuali conflitti tra conclusioni con i due metodi applicati. Come citato da Miller (2019), utilizzare un numero molto alto di lag, in questo caso 50, è una scelta molto insolita nelle applicazioni a serie storiche, ma nella nostra analisi non è una scelta così sbagliata quella di dare una dimensione maggiore a p e una certa distanza tra ogni singola osservazione. Infatti, con 50 lag si riesce a trovare un equilibrio tra i test che vengono applicati sia alla serie storica interpolata linearmente e sia a quella interpolata *a step*, evitando così eventuali problemi sorti in precedenza relativi, ad esempio, a delle conclusioni diverse ottenute applicando i due diversi metodi alla stessa serie storica.

Consideriamo inizialmente i dati relativi all'interpolazione lineare. La prima cosa che è importante far notare dalla tabella è che, utilizzando l'interpolazione lineare applicata alla serie originaria (con zero ritardi), sia $\hat{\psi}_T(1)$ che $\hat{\psi}_T(2)$, ossia le statistiche relative al test della traccia, ci portano a rifiutare l'ipotesi nulla di 'non cointegrazione', facendoci supporre quindi che le

nostre serie siano in qualche modo cointegrate. Questa scelta è supportata anche dalle due statistiche $\hat{\rho}_T$ e $\hat{\tau}_T$ e va contro ciò che avevamo dedotto dalle conclusioni relative alla tabella precedente, dove sia con $p = 0$ che con $p = 1$ i test effettuati sulle serie analizzate attraverso l'interpolazione lineare erano risultati non robusti, e quindi piuttosto distorti.

Dopo aver applicato il metodo alla serie originaria, l'analisi prosegue aumentando di un lag la serie, fino ad arrivare a un totale di 50 lag. Le statistiche $\hat{\psi}_T(2)$, $\hat{\rho}_T$ e $\hat{\tau}_T$ subiscono delle variazioni in termini numerici, ma la scelta relativa al test resta sempre la stessa, ossia rifiutare l'ipotesi nulla di 'non cointegrazione'. La statistica relativa al test della traccia con $j = 1$, ossia $\hat{\psi}_T(1)$, fa capolgere completamente la scelta relativa al test. Infatti, essa varia da 35,6 (con $p = 0$) a 4,5 (con $p = 1$) o 5,8 (con $p = 10$), e i due valori relativi a 1 e 10 lag risultano essere minori del valore nominale, quindi non si rifiuta più l'ipotesi nulla di 'non cointegrazione'.

Per quanto riguarda l'interpolazione *a step*, invece, la valutazione è un po' diversa. Infatti, come possiamo notare dalla tabella, i valori relativi alle statistiche del test della traccia ci portano a rifiutare sempre l'ipotesi nulla di 'non cointegrazione', a prescindere dal valore di p. Perciò, i risultati che otteniamo non ci portano a conclusioni utili per il nostro test.

Ci sono vari motivi che possono causare questo inconveniente. Tra questi, uno dei più probabili è la presenza di correlazione seriale a medio-lungo termine, come citato da Miller (2019), causata dalla tecnica dell'interpolazione che potrebbe aver 'nascosto' questo particolare fenomeno che crea qualche problema nei test statistici. Un altro fenomeno che potrebbe provocare l'inconveniente citato in precedenza può essere la presenza di 'outliers', i quali sono delle osservazioni isolate che si distinguono particolarmente dalle altre e che possono far cambiare la scelta relativa al rifiuto o non rifiuto dell'ipotesi nulla di un test della traccia. Un esempio molto frequente, nell'ambito dei dati paleoclimatici, può essere un dato raccolto tra un periodo freddo e uno caldo, il quale potrebbe essere molto diverso da tutti gli altri dati presenti e, quindi, rivelarsi 'pericoloso' ai fini dell'inferenza statistica.

Osservando la tabella, si può subito notare che utilizzando l'interpolazione lineare con la serie originaria (p = 0), il test statistico della traccia ci porta a rifiutare entrambe le ipotesi nulle, e ciò sta a significare che le serie risultano essere in qualche modo cointegrate tra di loro. Infatti, se rifiutiamo l'ipotesi nulla della presenza di j trend stocastici significa che ne

Tabella 2: Alcuni test statistici applicati a sistemi paleoclimatici bivariati, Miller (2019)

	p	$\hat{\psi}_T(1)$	$\hat{\psi}_T(2)$	$\hat{\rho}_T$	$\hat{\tau}_T$
$\hat{z}_t^L$	0	$35,6$	$16208,3$	$23,5$	$3,4$
	1	$4,5$	$479,4$	$11789,1$	$19,1$
	$\vdots$				
	10	$5,8$	$290,2$	$4367,0$	$13,7$
	$\vdots$				
	50	$14,1$	$305,6$	$356,6$	$15,1$
$\hat{z}_t^S$	0	$13,9$	$462,9$	$789,9$	$19,9$
	1	$13,9$	$463,5$	$791,3$	$19,9$
	$\vdots$				
	10	$13,9$	$459,5$	$789,7$	$19,8$
	$\vdots$				
	50	$14,1$	$305,6$	$356,6$	$15,1$

sono presenti in numero inferiore a j e, perciò, ci saranno delle relazioni di cointegrazione pari a 2 - j nel nostro caso.

7 Conclusioni

L'analisi riportata in questo articolo, descritta nell'articolo di Miller (2019), è stata svolta per capire la dimensione empirica e la potenza di un test di cointegrazione standard, effettuato su serie storiche costituite da dati che vengono osservati irregolarmente o non contemporaneamente. I metodi utilizzati per trattare queste particolari tipologie di serie sono l'interpolazione lineare e quella *a step*. In termini matematici, il primo metodo si differenzia dal secondo in quanto necessita di dati futuri, mentre l'interpolazione *a step* considera le osservazioni al tempo presente e quelle passate.

Il problema dell'irregolarità delle serie è quasi sempre presente quando si analizzano dati di tipo paleoclimatico, e per cercare di risolverlo si applicano diverse strategie. Nella nostra analisi si può notare come, utilizzando il metodo dell'interpolazione lineare, i test di cointegrazione producono dei risultati che sono distorti. Questa conclusione è supportata anche da Ghysels e Miller (2014) che, nel contesto di serie storiche a frequenza mista, analizzano dei test di cointegrazione applicati a dei dati interpolati linearmente e giungono alla conclusione che questo metodo causa una notevole distorsione nei risultati, e perciò non è adatto per trattare serie storiche di questo tipo.

L'interpolazione *a step*, invece, è un metodo che sembra funzionare correttamente quando viene applicato alle serie che vengono analizzate, e i risultati che si ottengono dai test di cointegrazione che vengono effettuati sono una conferma di ciò, in quanto il test risulta essere robusto a prescindere dal valore che assume p, ossia del numero di ritardi che vengono considerati.

I risultati che sono stati ottenuti attraverso le simulazioni di Monte Carlo hanno dimostrato come l'interpolazione *a step* sia un metodo che, applicato a serie paleoclimatiche e quindi irregolari, porta a dei risultati che dimostrano una buona dimensione empirica e un'ottima potenza del test di cointegrazione. Inoltre, Kaufmann e Juselius (2013) hanno dimostrato, attraverso l'utilizzo di un modello autoregressivo vettoriale cointegrato, che esiste una relazione di cointegrazione tra le serie oggetto di studio in questo articolo, ossia quella della temperatura e quella della concentrazione di anidride carbonica nell'aria, a dimostrazione che la distorsione dimensionale che si ottiene quando viene applicata l'interpolazinoe lineare può essere dovuta, a volte, a cause non conosciute.

L'analisi di questa tipologia di dati, ricavati da sezioni di ghiaccio, sono

molto interessanti da analizzare ma, soprattutto, importanti da approfondire, in quanto non solo si possono studiare i fenomeni metereologici del passato fino a tanti anni fa, ma si può anche cercare di prevedere cosa avverrà in futuro, ad esempio lo scioglimento di qualche ghiacciaio, l'innalzamento delle acque dei mari e degli oceani, il riscaldamento globale del pianeta e altri fenomeni che possono caratterizzare il futuro nostro e quello di tutto il pianeta in generale.

8 Ringraziamenti

Dopo aver raggiunto questo obiettivo vorrei, innanzitutto, ringraziare la mia relatrice, professoressa Luisa Bisaglia, che si è sempre resa disponibile per spiegazioni e chiarimenti riguardanti la tesi, nonostante la situazione da dicembre a questa parte non sia stata molto d'aiuto. Ringrazio anche tutti gli altri professori della triennale, in quanto mi hanno fatto crescere come persona e hanno migliorato la mia conoscenza in ambito statistico.

Oltre ai professori, le prime persone che mi sento in dovere di ringraziare sono mamma e nonna, che mi hanno accompagnato in questa avventura e che hanno creduto in me dal primo giorno che ho iniziato l'Università, ma anche Alessandra e Veronica M., due persone che mi sono state a fianco sempre in questo percorso, soprattutto negli ultimi 7/8 mesi caratterizzati da eventi che lo hanno reso un pò difficile. Non hanno mai smesso di credere in me e di spronarmi e se sono riuscito a raggiungere questo traguardo è anche per merito loro.

Ringrazio poi Kevin, Matteo, Veronica R. e tutti quelli della mia compagnia che nel momento opportuno hanno saputo come incentivarmi e, come loro, anche i compagni e le compagne di corso, tra cui Ambra, Simone, Lia, Alex, è stato un piacere aver portato a termine questo percorso con voi. Grazie a tutti.

Riferimenti bibliografici

[1] G Amisano e C Giannini. *Topics in Structural VAR Econometrics.* Springer Berlin Heidelberg, 2012.

[2] F Busetti e AMR Taylor. «Stationarity tests for irregularly spaced observations and the effects of sampling frequency on power». In: Econometric Theory 21 (2005), pp. 757–794.

[3] RF Engle e CWJ Granger. «Cointegration and Error Correction: Representation, Estimation and Testing». In: Econometrica 55 (1987), pp. 251–276.

[4] E Ghysels e JI Miller. *On the size distortion from linearly interpolating low-frequency series for cointegration tests. In* Essays in Honor of Peter C.B. Phillips (Advances in Econometrics). A cura di Chang Y (eds.) Fomby TB Park JY. Vol. 33. 93-122. Bingley, UK: Emerald Group Publishing Limited, 2014.

[5] E Ghysels e JI Miller. «Testing for cointegration with temporally aggregated and mixed-frequency time series». In: Journal of Time Series Analysis 36 (2015), pp. 797–816.

[6] CWJ Granger e AA Weiss. *Time series analysis of error-correction models. In* Studies in econometrics, Time Series and Multivariate Statistics. A cura di Leo A. Goodman Samuel Karlin Takeshi Amemiya. 255-278. Academic Press, 1983.

[7] S Johansen e K Juselius. «Maximum likelihood estimation and inference on cointegration - with applications to the demand for money». In: Oxford Bulletin of Economics and Statistics 52 (1990), pp. 169–210.

[8] Masson Delmotte V Jouzel J et al. «Orbital and millennial Antarctic climate variability over the past 800,000 years». In: Science 317 (2007), pp. 793–796.

[9] RK Kaufmann e K Juselius. «Testing about hyphoteses about glacial cycles against the observational record». In: Paleoceanography 28 (2013), pp. 175–184.

[10] Le Floch M Luthi D et al. «High-resolution carbon dioxide concentration record 650,000-800,000 years before present». In: Nature 453 (2008), pp. 379–382.

[11] JI Miller. «Testing Cointegrating Relationships Using Irregular and Non-Contemporaneous Series with an Application to Paleoclimate Data». In: Journal of Time Series Analysis 40.6 (2019), pp. 936–950.

[12] Beer J Parrenin F Barnola J-M et al. «The EDC3 chronology for the EPICA Dome C ice core». In: Climate of the Past 3 (2007), pp. 485–497.

[13] F Peracchi. *Dizionario di Economia e Finanza*. 2012. URL: `www.treccani.it/enciclopedia/cointegrazione_(Dizionario-di-Economia-e-Finanza)/`.

[14] PCB Phillips e S Ouliaris. «Asymptotic properties of residual-based tests for cointegration». In: Econometrica 58 (1990), pp. 165–193.

[15] H Riebeek. *Paleoclimatology: the Oxygen balance*. 2005. URL: `https://earthobservatory.nasa.gov/features/Paleoclimatology_OxygenBalance`.

[16] KF Ryan e DEA Giles. *Testing for unit roots in economic time-series with missing observations. In* Messy data (Advances in Econometrics). A cura di Hill RC (eds.) Fomby TB. Vol. 13. 203-242. Bingley, UK: Emerald Group Publishing Limited, 1999.

[17] MW Watson. «Vector autoregressions and cointegration». In: Handbook of Econometrics 4 (1994). A cura di Elsevier, pp. 2843–2915.

[18] N Zingarelli. *Vocabolario della lingua italiana*. A cura di Zanichelli. P.1571. 2010.