UNIVERSITÀ POLITECNICA DELLE MARCHE
FACOLTÀ DI ECONOMIA "GIORGIO FUÀ"

Corso di Laurea Magistrale o Specialistica in Scienze Economiche e Finanziarie

"ANALISI E PREVISIONE DEL MERCATO DELLE CRIPTOVALUTE: MODELLI PARAMETRICI E NON PARAMETRICI"

"CRYPTOCURRENCY MARKET ANALYSIS AND FORECASTING: PARAMETRIC AND NON-PARAMETRIC MODELS"

Relatore: Chiar.ma
Prof.ssa Francesca Mariani

Tesi di Laurea di:
Federico Olivi

Anno Accademico 2019 – 2020

INDICE

INTRODUZIONE

In questa tesi verrà trattato il tema dei big data riferiti al social media Twitter e come questi possano in qualche modo influenzare il mercato delle criptovalute. Siti di microblogging come Twitter si sono infatti evoluti diventando una fonte di diversi tipi di informazione. Gli individui pubblicano su questi social media messaggi in tempo reale circa le loro opinioni su una varietà di soggetti, discutono di eventi futuri, si lamentano ed esprimono giudizi positivi per i prodotti utilizzati nella vita quotidiana. Tali piattaforme di scambio di messaggi hanno pertanto acquisito importanza per le aziende che vendono prodotti o servizi, ricevendo un feedback generale circa la soddisfazione del consumatore finale.

Quello che si è cercato di fare in questa tesi, in seguito alla lettura dell'articolo "The predictive power of public Twitter sentiment for forecasting cryptocurrency prices", è di utilizzare lo stesso approccio adottato dalle aziende per ricercare un feedback sui loro prodotti. Infatti, con il fine ultimo di avere un quadro generale più ampio circa il futuro andamento delle criptovalute considerate, si è cercato di andare ad analizzare le opinioni, le news e le informazioni che circolano tra gli individui sul social media Twitter in relazione alle valute digitali.

Le criptovalute sono valute digitali che fanno uso della tecnologia blockchain, una tecnologia innovativa, decentralizzata e crittografica che consente la

digitalizzazione della fiducia tra diversi operatori. In particolare, nel contesto delle criptovalute la tecnologia blockchain svolge il ruolo dei governi come produttori di valuta e il ruolo degli intermediari nel verificare che una determinata transazione sia diventata obsoleta. Sebbene le criptovalute siano nate in seguito al lancio di Bitcoin nel 2009, il loro dirompente potenziale che ha portato ad una crescita esponenziale dell'interesse nelle valute digitali si è affermato nel corso del 2017 e all'inizio del 2018. Proprio le notizie circa i rendimenti senza precedenti delle criptovalute hanno attirato l'attenzione su quest'ultime, incrementando l'interesse degli operatori finanziari inducendoli a volersene impossessare. Nonostante ciò, la normativa riguardante le criptovalute risulta incompleta non essendo queste ancora riconosciute come delle classi di asset maturi. Proprio le lacune normative, combinate all'alta popolarità e alla mancanza di un'istituzione che svolge un ruolo di garante hanno reso il mercato delle criptovalute altamente volatile.

In questa tesi, dovutamente all'alta volatilità delle criptovalute, è stato studiato il loro comportamento all'interno della singola giornata di negoziazione considerando i prezzi e i volumi battuti ogni minuto. Per studiare il loro andamento futuro sono stati adottati sia modelli parametrici che non parametrici. In particolare, sono state svolte analisi per indagare circa il potere previsionale delle opinioni e informazioni estratte da Twitter sulle criptovalute considerate mediante modelli parametrici. Inoltre, quest'ultimi sono stati utilizzati per effettuare previsioni delle variabili finanziarie relative alla criptovalute tenendo in considerazione le informazioni di

Twitter. Successivamente, le previsioni ottenute attraverso il modello parametrico adottato sono state aggiustate mediante il modello non parametrico ottenendo quindi una previsione alternativa.

La seguente tesi è articolata in tre capitoli, in cui i primi due capitoli trattano e descrivono gli strumenti e i processi utilizzati per ottenere i risultati illustrati nel terzo capitolo.

Nel primo capitolo vengono illustrati i concetti teorici dei modelli parametrici e non parametrici utilizzati nello svolgimento delle analisi. Nello specifico, viene descritto il modello vettoriale autoregressivo VAR andando a sottolineare le principali considerazioni teoriche annesse, i suoi utilizzi, i vantaggi e gli svantaggi connessi alla sua adozione. In seguito, verrà illustrato il tema relativo alle tecniche di Machine Learning non parametriche. In particolare, l'attenzione sarà concentrata sul classificatore K-Nearest-Neighbor, il quale verrà utilizzato per aggiustare le previsioni effettuate con il VAR.

Il secondo capitolo sarà incentrato principalmente nella descrizione del mercato delle criptovalute e dei fattori che determinano la sua prevedibilità. Inoltre, verranno discusse le potenzialità di Twitter nel prevedere i movimenti, andando a descrivere come sono stati ottenuti e convertiti in punteggio i tweet relativi alle criptovalute considerate. In particolare, verrà descritta l'importanza della Sentiment Analysis e delle tecniche di pre-processamento adottate per attribuire ai tweet raccolti un valore numerico.

Nell'ultima parte del secondo capitolo verranno descritti i processi e le metodologie utilizzate per svolgere l'analisi di causalità e il miglioramento delle previsioni del VAR.

Infine, nel terzo capitolo verranno discussi e confrontati i risultati delle analisi effettuate in relazione alle criptovalute e alle diverse frequenze di osservazioni valutate.

Capitolo 1

MODELLI PARAMETRICI E NON PARAMETRICI

1.1 MODELLO VETTORIALE AUTOREGRESSIVO VAR(*p*)

Un investitore, o più in generale un responsabile delle decisioni, nell'intraprendere una scelta a livello strutturale necessita spesso di previsioni di variabili economiche. Se le osservazioni attuali e passate della variabile d'interesse sono disponibili e contengono importanti informazioni previsionali sullo sviluppo futuro di una variabile, risulta convenevole utilizzare come previsione una funzione dei dati osservati nel passato.

Tuttavia, è piuttosto raro che in ambito economico e finanziario un fenomeno possa essere descritto solamente da una variabile e dalle sue osservazioni passate, in aggiunta, questo potrebbe essere descritto anche da altre variabili e dalle relative osservazioni passate. Anche in questo caso, in analogia con il caso in cui ci si interessi ad una sola variabile, l'obiettivo è quello di determinare una funzione che possa essere utilizzata per ottenere buone previsioni delle variabili prese in considerazione.

In questa tesi verrà utilizzato un approccio multivariato all'analisi delle serie storiche basato sul modello vettoriale autoregressivo VAR (*p*).

I modelli vettoriali autoregressivi VAR sono stati introdotti da Sims (1980) come metodo alternativo a quelli utilizzati nelle analisi macroeconomiche classiche, ad esempio le equazioni simultanee[1], le quali presentano problemi di identificazione. Infatti, la principale critica di Sims a questi modelli macroeconometrici è basata sul fatto che questi non poggino su solide teorie economiche o che le teorie economiche non siano in grado di fornire una specificazione completa del modello.

Concettualmente i modelli vettoriali autoregressivi VAR sono la generalizzazione multivariata dei modelli autoregressivi AR[2], in cui il valore al tempo *t* della variabile in esame è funzione lineare dei propri valori passati più un *white noise*. Analogamente, nei modelli VAR il valore al tempo *t* di ciascuna delle variabili in esame è descritto da una funzione lineare delle proprie osservazioni passate e delle osservazioni passate delle altre variabili più un *white noise.*

In particolare, un modello vettoriale autoregressivo di ordine $p \geq 1$ VAR (p) con $K > 1$ variabili si presenta nella seguente forma:

$$y_t = v + A_1 y_{t-1} + \cdots + A_p y_{t-p} + u_t\,, \qquad t = 0, \pm 1, \pm 2, \pm 3, \ldots, \tag{1.1}$$

dove $y_t = (y_{1t}, \ldots, y_{Kt})'$ è un vettore casuale $(K \times 1)$, $A_i, i = 1,2, \ldots, p,$ sono matrici di coefficienti $(K \times K)$, $v = (v_1, \ldots, v_K)'$ è il vettore i termini di intercetta la cui presenza permette a y_t di avere media diversa da zero, ed infine il vettore di *white noise* $u_t = (u_{1t}, \ldots, u_{Kt})'$ con dimensione $(K \times 1)$ le cui medie e covarianze

[1] *Sims, "Macroeconomics and Reality", Econometrica, Vol. 48, No. 1, pp. 1-48.*
[2] *Gallo, Pacini, "Metodi Quantitativi per i Mercati Finanziari", Carocci, pp. 195.*

sono pari, rispettivamente, a $E(u_t) = 0$, $E(u_t u_t') = \Sigma_u$ e $E(u_t u_s') = 0$ per ogni $s \neq t$, dove Σ_u è una matrice $(K \times K)$.

1.1.1 Stabilità e stazionarietà dei processi VAR(p)

Un aspetto cruciale dei modelli vettoriali autoregressivi VAR(p) è quello di indagare circa la stazionarietà del processo analizzato, poiché quest'ultima è una condizione necessaria purché le stime dei parametri inseriti nel modello risultino significative. In generale, un processo stocastico[3] è stazionario se il suo momento di ordine primo e il suo momento di ordine secondo non variano al variare del tempo *t.*

Con l'obiettivo di analizzare la stazionarietà di un processo VAR(p) si prenda in considerazione un processo del primo ordine VAR ($p=1$):

$$y_t = v + A_1 y_{t-1} + u_t, \qquad t = 1,2, \dots, \tag{1.2}$$

Se il processo generatore della serie storica inizia al tempo 1, allora la dinamica del processo y_tè descritta dalle equazioni:

$$y_1 = v + A_1 y_0 + u_1$$

$$\vdots$$

$$y_t = (I_K + A_1 + \cdots + A_1^{t-1})v + A_1^t y_0 + \sum_{i=0}^{t-1} A_1^i u_{t-i}, \qquad t = 1,2, \dots, \tag{1.3}$$

Dove I_K è la matrice identità di ordine *K*.

[3] Un processo stocastico è una famiglia di variabili casuali definito in uno spazio probabilistico. *Brockwell, Davis,"Time Series: Theory and Methods" Second Edition, Springer, pp. 8.*

Quindi i vettori $y_1, \dots, y_t$ e le loro distribuzioni congiunte sono univocamente determinati da y_0 e $u_1, \dots, u_t$ e dalle loro distribuzioni congiunte.

Tuttavia, spesso è conveniente assumere che il processo generatore sia iniziato in un generico istante nel passato $t - j, j = 1,2, \dots, t - 1$, in questo caso il processo VAR (1) può essere riscritto come segue:

$$y_t = v + A_1 y_{t-1} + u_t =$$

$$= \left(I_K + A_1 + \cdots + A_1^j\right)v + A_1^{j+1} y_{t-j-1} + \sum_{i=0}^{j} A_1^i \, u_{t-i}, t = 1,2, \dots \ . \qquad (1.4)$$

Se tutti gli autovalori di A_1 sono in modulo inferiori ad 1, la sequenza A_1^i risulta essere sommabile. Quindi, la somma infinita $\sum_{i=0}^{\infty} A_1^i \, u_{t-i}$ esiste in media quadrata e il termine $A_1^{j+1} y_{t-j-1}$ può essere ignorato nel limite poiché A_1^{j+1} converge rapidamente a zero per $j \to \infty$[4]. Se si rispetta la condizione per cui gli autovalori di A_1 risultano in modulo inferiori ad 1 si può affermare che y_t è un processo stocastico ben definito:

$$y_t = \mu + \sum_{i=0}^{\infty} A_1^i \, u_{t-i}, \quad \mu = (I_K - A_1)^{-1} v, \qquad t = 0, \pm 1, \pm 2, \pm 3 \dots, \qquad (1.5)$$

In quest'ultima forma, la distribuzione e la distribuzione congiunta di y_t sono unicamente determinate dalle distribuzioni dei processi *white noise* u_t, e il

[4] $\left(I_K + A_1 + \cdots + A_1^j\right)v \xrightarrow[j \to \infty]{} (I_K - A_1)^{-1} v$

momento primo e secondo di tale processo stocastico possono essere definiti come segue:

- $E(y_t) = \mu$, per ogni *t.*
- $\Gamma_y(h) = \sum_{i=0}^{\infty} A_1^{h+i} \Sigma_u A_1^{i\prime}$, per ogni *t*.

Dunque, data l'importanza della condizione posta sugli autovalori di A_1, un processo VAR (*1*) può essere definito stabile se gli autovalori di A_1 risultino in modulo essere inferiori ad 1. La condizione di stabilità può essere riscritta in seguito a dei passaggi algebrici nella seguente forma:

$$\det(I_K - A_1 z) \neq 0 \qquad per\ |z| \leq 1 \tag{1.6}$$

La precedente analisi potrebbe essere facilmente estesa ai processi VAR di ordine *p* poiché ogni processo VAR(*p*) può sempre essere riscritto nella forma VAR (*1*) attraverso l'utilizzo della rappresentazione in *companion form*[5].

Perciò, l'estensione della condizione di stabilità per un processo VAR di ordine *p* può essere rappresentata dalla seguente condizione:

$$\det\left(I_K - A_1 z - \cdots - A_p z^p\right) \neq 0 \qquad per\ |z| \leq 1 \tag{1.7}$$

Dato che la stabilità implica la stazionarietà, nella letteratura delle serie storiche spesso ci si riferisce alla condizione di stazionarietà riferendosi alle condizioni sopra riportate. In conclusione, un processo stabile VAR(*p*), y_t, per $t = 0, \pm 1, \pm 2, \pm 3 \ldots$, è stazionario.

[5] *Ooms, "Empirical Vector Autoregression Model", pp. 234.*

1.1.2 Stima dei processi VAR(p)

La stazionarietà di un processo VAR(p) è condizione necessaria poiché le stime dei parametri effettuate con il metodo dei minimi quadrati ordinari (OLS)[6] siano consistenti. Infatti, se si desidera utilizzare test d'ipotesi, singolarmente o congiuntamente, per analizzare la significatività dei coefficienti stimati, è necessario che tutte le serie inserite nel modello VAR(p) siano stazionarie. Si consideri un VAR di ordine *p* n-variato rappresentato nella seguente forma:

$$y_{it} = \sum_{j=1}^{p}\left(a_{i1j}y_{1t-j} + \cdots + a_{in}y_{nt-j}\right) + u_{it}\,, t = 1,2,\dots, i = 1,2,\dots,n. \quad (1.8)$$

Ognuna delle *n* equazioni presenti nel processo VAR(p) sopra illustrato può essere vista come un modello di regressione dinamica, e si possono produrre stime consistenti e asintoticamente normali dei coefficienti $a_{ij}, i = 1,2,\dots,n,\ j = 1,2,\dots,p$, attraverso l'applicazione dei minimi quadrati ordinari (OLS).

Relativamente alla stima della matrice delle varianze-covarianze Σ_u, l'applicazione del metodo OLS alle *n* equazioni produce *n* serie di residui $\widehat{u_1},\dots\widehat{u_n}$ e si può mostrare che la covarianza campionaria tra i residui è uno stimatore consistente in senso debole della componente *st* della matrice varianze-covarianze Σ_u.

$$\frac{1}{T}\widehat{u_t}\widehat{u_s}' \xrightarrow{p} \Sigma_u \quad (1.9)$$

Infatti, gli stimatori ottenuti mediante l'applicazione del metodo dei minimi quadrati ordinari (OLS) risultano essere corretti e Best Linear Unbiased Estimator

[6] *Hamilton, "Time Series Analysis", pp. 200.*

(*BLUE*) in campioni finiti, mentre godono della proprietà della consistenza e della distribuzione asintotica normale nel caso di campioni infiniti.

1.1.3 Previsione con i VAR(p)

Uno dei possibili utilizzi dei processi vettoriali autoregressivi VAR(p) è quello di ottenere previsioni nel futuro sfruttando le informazioni contenute nei dati delle serie storiche a disposizione. Infatti, l'investitore solitamente si trova nella situazione in cui ad un determinato periodo temporale *t* deve fare supposizioni circa il valore futuro di variabili $y_{t+1}, \dots, y_{t+K}$. Per farlo, utilizza un processo generatore dei dati, in questo caso un processo VAR(p), avendo l'obiettivo di minimizzare una specifica funzione di costo o di perdita associata all'errore della previsione.

Nel contesto dei modelli VAR, i previsori che minimizzano la somma dei quadrati degli errori (Means Squared Errors (*MSE*)) commessi approssimando la serie storica considerata con il VAR sono ampiamenti i più utilizzati. Argomentazioni sull'uso dei MSE per definire una funzione di perdita sono state esposte da Granger (1969), e Granger & Newbold (1986), in particolare, gli autori dimostrano che la previsione che rende minima la funzione di perdita attesa definita attraverso gli MSE minimizza anche altre funzioni di perdita attesa.

Concettualmente le previsioni con il VAR(p) possono essere di due tipologie: puntuali o intervallari. Quest'ultima tipologia di previsione richiede delle assunzioni di base circa le distribuzioni di y_t o u_t, i quali devono essere considerati

processi Gaussiani. Tuttavia, in questo elaborato verrà trattata esclusivamente la previsione di tipo puntuale mediante i processi vettoriali autoregressivi VAR(*p*).

Si supponga che $y_t = (y_{1t}, \dots, y_{Kt})'$ sia un processo VAR(p) stabile e stazionario K-dimensionale.

Quindi, il previsore MSE minimo per un orizzonte di previsione *h* all'origine della previsione *t* è il seguente valore atteso condizionale:

$$E_t(y_{t+h}) = E(y_{t+h}|\Omega_t) = E(y_{t+h}|\{y_s|s \leq t\}) \tag{1.10}$$

Il previsore minimizza il MSE di ogni componente di y_t. Dunque, se $\overline{y_t}(h)$ è un qualunque previsore al tempo $t + h$, si avrà che:

$$MSE[\overline{y_t}(h)] = E[\big(y_{t+h} - \overline{y_t}(h)\big)\big(y_{t+h} - \overline{y_t}(h)\big)']$$

$$\geq MSE[E_t(y_{t+h})] = E\left[\big(y_{t+h} - E_t(y_{t+h})\big)\big(y_{t+h} - E_t(y_{t+h})\big)'\right] \tag{1.11}$$

Dove il segno di disuguaglianza tra le due matrici sta ad indicare che la differenza tra il termine a sinistra e il termine a destra è una matrice semidefinita positiva.

L'ottimalità della previsione ottenuta mediante il valore atteso condizionale può essere sottolineata notando che:

$$\begin{aligned} MSE[\overline{y_t}(h)] &= E\big[[y_{t+h} - E_t(y_{t+h}) + E_t(y_{t+h}) \\ &\quad - \overline{y_t}(h)][y_{t+h} - E_t(y_{t+h}) + E_t(y_{t+h}) - \overline{y_t}(h)]'\big] \\ &\quad = MSE[E_t(y_{t+h})] \\ &\quad + E\big[[E_t(y_{t+h}) - \overline{y_t}(h)][E_t(y_{t+h}) - \overline{y_t}(h)]'\big] \end{aligned} \tag{1.12}$$

Quest'ultimo risultato è valido poiché $\left(y_{t+h} - E_t(y_{t+h})\right)$ è una funzione delle innovazioni dopo il periodo *t* la quale risulta incorrelata con i termini contenuti in $(E_t(y_{t+h}) - \bar{y}_t(h))$ perché è funzione di $y_s, s \leq t$. L'ottimalità della previsione ottenuta implica che:

$$E_t(y_{t+h}) = v + A_1 E_t(y_{t+h-1}) + \cdots + A_p E_t\left(y_{t+h-p}\right),\ t = 1,2,\dots,h > 0, \quad (1.13)$$

Pertanto, y_{t+h} risulta essere il previsore ottimo al tempo *h* di un processo VAR(*p*) a patto che u_t sia un processo di tipo *white noise* indipendente così che u_t e u_s siano indipendenti per $s \neq t$ e $E_t(u_{t+h}) = 0$ per $t = 1,2,\dots,h > 0$.

Le previsioni ottenute mediante il metodo dei valori attesi condizionati possiedono le seguenti proprietà:

- sono previsioni corrette, in quanto $E_t[y_{t+h} - E_t(y_{t+h})] = 0$,
- se u_t risulta essere un *white noise* indipendente, $MSE[E_t(y_{t+h})] = MSE[E_t(y_{t+h})|y_t, y_{t-1}, \dots]$.

1.1.4 Analisi di causalità

Un'altra applicazione per cui i processi vettoriali autoregressivi VAR(*p*) sono comunemente utilizzati è l'analisi della causalità. Infatti, poiché i modelli VAR rappresentano dei sistemi di variabili che lo compongono, questi modelli vengono spesso utilizzati per indagare alcuni aspetti tra le relazioni delle variabili d'interesse.

In generale, le relazioni di causa-effetto sono molto complesse da stabilire in ambito economico, perché se osserviamo un'alta correlazione tra due variabili, si può concludere al massimo che queste tendono a muoversi insieme, ma, in assenza di altre informazioni specifiche, non sarà possibile concludere circa gli ipotetici nessi causali tra le due.

Granger (1969) ha definito un concetto di causalità il quale, sotto determinate condizioni, è facilmente osservabile nel contesto dei processi VAR. L'idea che sta alla base del concetto di causalità secondo Granger è che la causa non può avvenire successivamente all'effetto. Quindi, se una variabile *x* influenza una variabile *z*, la prima deve migliorare la previsione effettuata della seconda variabile. Per formalizzare questo concetto, assumiamo che $\mathrm{z}_t(h|\Omega_t)$ sia il previsore ottimale del processo z_t al tempo $\mathrm{t}+\mathrm{h}$, basato sulle informazioni contenute in Ω_t. La previsione MSE corrispondente effettuata mediante il previsore $\mathrm{z}_t(h|\Omega_t)$ sarà rappresentata da $\Sigma_z(h|\Omega_t)$. Il processo x_t si dice "granger causa" z_t se:

$$\Sigma_z(h|\Omega_t) < \Sigma_z(h|\Omega_t \backslash \{x_s|s \leq t\}) \quad per\ almeno\ un\ h = 1,2,\ldots,$$
$$t = 1,2,\ldots \tag{1.14}$$

Nell'Equazione 1.14 il termine $\Omega_t \backslash \{x_s|s \leq t\}$ sta ad indicare che le informazioni considerate sono tutte quelle a disposizione tranne le informazioni contenute nel passato e nel presente del processo x_t. Perciò, se z_t può essere previsto più efficacemente se le informazioni contenute nel processo x_t sono prese in considerazione in aggiunta alle altre informazioni, x_t granger causa z_t. Se, x_t

granger causa z_t e z_t granger causa x_t il processo $(z_t' x_t')'$ viene denominato sistema *feedback*.

Per indagare circa la relazione di Granger-causalità tra le variabili che compongono un processo VAR K-dimensionale y_t, si prenda in considerazione la rappresentazione a media mobile[7] dello stesso processo:

$$y_t = \mu + \sum_{i=0}^{\infty} \phi_i u_{t-i} = \mu + \phi(L)u_t, \quad t = 1,2,\dots, \quad \phi_0 = I_K \tag{1.15}$$

Supponendo che y_t sia partizionato in un processo M-dimensionale z_t e un processo (K-M)-dimensionale x_t, usando la formula previsionale, la previsione ottimale al periodo 1 di z_t basato su y_t risulta essere:

$$z_t(1|\{y_s|s \le t\}) = [I_M \quad \underline{0}]y_t(1) = \mu_1 + \sum_{i=1}^{\infty} \phi_{11,i} u_{1,t+1-i} + \sum_{i=1}^{\infty} \phi_{12,i} u_{2,t+1-i} \tag{1.16}$$

Quindi l'errore di previsione è dato da:

$$z_{t+1} - z_t(1|\{y_s|s \le t\}) = u_{1,t+1} \tag{1.17}$$

Essendo quest'ultimo un sotto-processo di un processo stazionario, anch'esso ha una rappresentazione in media mobile dell'errore di previsione:

$$z_t = \mu_1 + \sum_{i=1}^{\infty} \phi_{11,i} u_{1,t-i} + \sum_{i=1}^{\infty} \phi_{12,i} u_{2,t-i} = \mu_1 + \sum_{i=1}^{\infty} F_i v_{t-i}, t = 1,2,\dots \tag{1.18}$$

[7] *Lütkepohl, "New Introduction to Multiple Time Series Analysis", Springer, pp.18.*

dove $F_0 = I_M$ e l'ultima espressione descrive la rappresentazione in media mobile dell'errore di previsione.

Quindi il previsore ottimale ad 1 periodo basato solamente su z_t e il corrispondente errore di previsione risultano essere:

$$z_t(1|\{z_s|s \leq t\}) = \mu_1 + \sum_{i=1}^{\infty} F_i v_{t+1-i}\,, t = 1,2, \ldots,$$

$$z_{t+1} - z_t(1|\{z_s|s \leq t\}) = v_{t+1}\; t = 1,2, \ldots, \tag{1.19}$$

Di conseguenza i previsori di z_t, uno basato su y_t e l'altro basato solamente su z_t, saranno identici se e solo se $v_t = u_{1,t}$ per ogni *t,* $t = 1,2, \ldots,$. In conclusione, essendo y_t un processo VAR con l'operatore a media mobile $\phi(L)$, allora:

$$z_t(1|\{y_s|s \leq t\}) = z_t(1|\{z_s|s \leq t\}) \iff \phi_{12,i} = 0$$
$$per\ i = 1,2, \ldots, t = 1,2, \ldots, \tag{1.20}$$

Quest'ultimo risultato è considerato come la condizione necessaria e sufficiente per affermare che il processo x_t non Granger-causa il processo z_t. Quindi, la non Granger-causalità può essere facilmente individuata andando ad analizzare la rappresentazione in media mobile del processo y_t.Tuttavia, essendo, in questa sede, interessati in particolare all'analisi della causalità in un processo vettoriale autoregressivo VAR(*p*) stabile e stazionario, la condizione di non Granger-causalità è soddisfatta se e solo se:

$$A_{12,i} = 0 \qquad per\ i = 1, \ldots, p. \tag{1.21}$$

La semplicità di applicazione dell'analisi di causalità secondo Granger ha portato ad una notevole diffusione della stessa nella letteratura econometrica e ha sollevato due principali critiche.

La prima critica, di carattere statistico, si concentra sul fatto che una variabile X può essere o meno Granger-causale per un'altra variabile Y a seconda di quali altre variabili siano presenti nel modello VAR. Ad esempio, in termini di Granger-causalità, un test che accetti l'assenza di Granger-causalità in un VAR bivariato potrebbe rifiutarla in un VAR trivariato, e quindi i test devono sempre essere considerati validi all'interno del set di condizionamento scelto.

La seconda critica, invece, fa riferimento al fatto che il concetto logico di causa-effetto prescinde da ciò che accade nel tempo fisico. In particolare, è possibile che la causa si manifesti solo dopo l'effetto, quando quest'ultimo è influenzato dalle aspettative degli agenti.

1.1.5 Scelta dell'ordine *p* per un processo VAR

Fino a questo punto sono stati considerati processi VAR(p) senza tenere in considerazione quale sia l'ordine p ottimale per far sì che il modello si adatti efficacemente alle serie storiche che lo compongono. Tuttavia, un modello VAR viene spesso costruito per effettuare previsioni delle variabili coinvolte, quindi l'obiettivo della ricerca dell'ordine dei ritardi p ottimale ha come finalità ultima

quella di individuare un efficace modello previsionale adattabile ai dati a disposizione.

Spesso, la teoria finanziaria, a proposito di quale sia la lunghezza dei ritardi da adottare in un processo VAR, risulta scarsa, per tale motivo possono essere introdotti solo due differenti metodi di scelta che possono essere utilizzati per determinare l'ordine *p* ottimale del processo VAR: le equazioni con restrizioni incrociate e i criteri d'informazione.

1.1.5.1 Equazioni con restrizioni incrociate

Per determinare l'ordine dei ritardi ottimali da inserire in un processo VAR, potrebbero, in prima battuta, essere utilizzati i test-F. Tuttavia, quest'ultimi risultano inappropriati in questa trattazione in quanto verrebbero utilizzati separatamente per l'insieme dei ritardi in ciascuna equazione. Invece, quello che viene richiesto per determinare il ritardo ottimale è una procedura per testare i coefficienti su un insieme di ritardi, su tutte le variabili per tutte le equazioni che compongono il VAR allo stesso tempo.

In questa sede, si può sottolineare come nello spirito della stima del VAR ai modelli debbano essere applicate meno limitazioni possibili, per questo motivo un VAR con diverse lunghezze di ritardo potrebbe essere descritto come un VAR limitato.

Un approccio alternativo applicabile potrebbe essere quello di specificare lo stesso numero di ritardi in ciascuna equazione e determinare l'ordine ottimale del modello

attraverso un test del rapporto di verosimiglianza. Indicando la matrice delle varianze-covarianze dei residui come $\widehat{\Sigma_u}$, il test del rapporto di verosimiglianza o likelihood ratio (LR) sotto queste ipotesi congiunte risulta essere:

$$LM = T\left[\log\left|\widehat{\Sigma_r}\right| - \log\left|\widehat{\Sigma_u}\right|\right] \tag{1.22}$$

dove $\left|\widehat{\Sigma_r}\right|$ è il determinante della matrice delle varianze-covarianze dei residui del modello limitato, $\left|\widehat{\Sigma_u}\right|$ è il determinante della matrice delle varianze-covarianze dei residui del modello non limitato e *T* è la dimensione del campione.

La statistica test LM è distribuita come una Chi-quadro (χ^2) con gradi di libertà pari al numero di restrizioni imposte. Nel caso generale di un VAR con *n* equazioni, per imporre la restrizione che gli ultimi *g* ritardi abbiano coefficienti pari a zero, debbono essere imposte in totale $n^2 g$ restrizioni.

1.1.5.2 Criteri d'informazione per la scelta dell'ordine dei ritardi

Il test del rapporto di verosimiglianza descritto in precedenza risulta intuitivo e di facile applicazione pur presentando alcune limitazioni. Principalmente, uno dei due VAR analizzati sopra deve essere un caso particolare dell'altro, e più specificatamente può essere effettuata solamente una comparazione tra due differenti modelli. Un ulteriore svantaggio del test del rapporto di verosimiglianza *LR* è che il test χ^2 sarà strettamente significativo asintoticamente solo se si considera l'ipotesi che i residui di ciascuna equazione siano distribuiti in modo

normale, la quale ipotesi risulta improbabile se si considera l'utilizzo di dati finanziari.

Un approccio alternativo al test del rapporto di verosimiglianza propone di utilizzare i criteri d'informazione per ricercare l'ordine ottimale dei ritardi *p.* Infatti, tali criteri d'informazione non richiedono alcuna ipotesi intrinseca di normalità relativamente alla distribuzione dei residui. Invece, un aspetto da sottolineare nell'approccio dei criteri d'informazione è che quest'ultimi compensano una diminuzione della somma dei quadrati dei residui (*RSS*) di ciascuna equazione man mano che vengono aggiunti più ritardi, attraverso l'incremento del termine di penalità. Per determinare l'ordine del processo ottimo potrebbero essere applicati i criteri d'informazione univariati separatamente a ciascuna equazione, tuttavia è solitamente preferibile individuare un numero di ritardi che sia lo stesso per ogni equazione inserita nel sistema VAR.

Per questo motivo, di seguito verranno illustrati tre differenti criteri d'informazione nella versione multivariata che possono essere descritti come:

$$Criterio\ d'informazione\ d'Akaike(AIK) = \log\left|\widehat{\Sigma_u}\right| + \frac{2k'}{T} \qquad (1.23)$$

$$Criterio\ d'informazione\ di\ Schwarz(SC) = \log\left|\widehat{\Sigma_u}\right| + \frac{k'\log(T)}{T} \qquad (1.24)$$

$$Criterio\ d'informazione\ di\ Hannan\ Quinn(HQ) = \log\left|\widehat{\Sigma_u}\right| + \frac{k'\log(\log(T))}{T} \qquad (1.25)$$

dove $\widehat{\Sigma_u}$ è la matrice delle varianze-covarianze dei residui, T è il numero delle osservazioni e k' è il numero dei regressori totali in tutte le equazioni, il quale risulta essere uguale a $n^2p + n$, dove n è il numero di equazioni inserite nel sistema VAR, ognuna con p ritardi delle n variabili, più una costante per ognuna delle n equazioni. I criteri d'informazione vengono costruiti per $0,1, \dots, \bar{p}$ ritardi, dove $\bar{p}$ è il valore del ritardo massimo pre-specificato, e si va a scegliere il numero dei ritardi che minimizza il valore dei criteri d'informazione sopra citati. In generale, andando a calcolare tutti e tre i criteri d'informazione, *AIK*, *SC* e *HQ*, la scelta dell'ordine del ritardo ottimale non è univoca. Infatti, per come sono costruiti i criteri d'informazione si ha di solito che l'*AIK* tende ad essere piuttosto permissivo in termini di numero totale di parametri, mentre il *BIC* solitamente ha la tendenza opposta e invece il criterio d'informazione *HQ* in genere opta per una soluzione intermedia. Tuttavia, il criterio d'informazione di Akaike risulta essere il più diffuso per la sua importanza storica e per essere stato il primo ad essere costruito rispetto agli altri criteri, nonostante ciò, esso conserva il difetto intrinseco di non essere consistente, in quanto può essere dimostrato che la probabilità che esso selezioni il modello corretto non tende ad 1 asintoticamente.

1.1.6 Vantaggi e svantaggi dei modelli VAR

In seguito alla trattazione delle principali questioni relative ai modelli vettoriali autoregressivi VAR(p) è possibile andare ad individuare e descrivere quali sono i

maggiori vantaggi e svantaggi nell'applicazione dei processi considerati in precedenza.

Infatti, i modelli VAR possiedono molteplici vantaggi se comparati ai modelli di serie storiche univariati o ai modelli strutturali di equazioni simultanee.

Un primo vantaggio nell'utilizzo dei modelli vettoriali autoregressivi è quello di non dover specificare quali delle variabili inserite nel processo VAR debba essere endogena oppure esogena, poiché tutte le variabili considerate all'interno del sistema risultano essere endogene. Questo punto risulta essere molto importante, dato che i modelli strutturali di equazioni simultanee richiedono, per essere stimati, che tutte le equazioni del sistema siano identificate. Tale requisito si riduce alla condizione che alcune variabili siano identificate come esogene e le equazioni del sistema contengano una diversità nelle variabili considerate. Idealmente, questa restrizione dovrebbe emergere dalla teoria economico-finanziaria ma, in pratica la teoria risulta essere apparentemente vaga circa quali delle variabili considerate debbano essere trattate come esogene. Di conseguenza viene lasciata un'elevata discrezionalità nella scelta su come classificare le variabili. In conclusione, la specificazione di variabili come esogene richiede delle forme di restrizione che sono state criticate da Sims[8] come "incredibili", motivo per cui la stima dei processi VAR non richiede che tali restrizioni debbano essere imposte.

[8] *Sims, "Macroeconomics and Reality", Econometrica, Vol. 48, No. 1, pp. 1-48.*

Un secondo vantaggio dei modelli VAR è che quest'ultimi permettono al valore di una variabile di dipendere da altre variabili in aggiunta ai propri ritardi e alla combinazione di termini residuali di tipo *white noise*. I processi VAR, pertanto, possono offrire una struttura maggiormente ricca e flessibile, essendo capaci di catturare maggiori caratteristiche dai dati disponibili, rispetto ad un modello autoregressivo univariato AR, il quale può essere considerato una forma limitata di un processo VAR.

Un ulteriore vantaggio dei modelli vettoriali autoregressivi è dato dal fatto che non essendoci termini contemporanei nelle equazioni specificate all'interno del sistema, la stima dei parametri può essere semplicemente effettuata applicando gli OLS separatamente ad ogni singola equazione.

Infine, come ultimo vantaggio va annoverato il fatto che le previsioni ottenute attraverso i processi VAR risultano migliori rispetto ai modelli strutturali tradizionali. Questo aspetto è stato argomentato in molteplici articoli (ad esempio Sims, 1980) in cui si osserva che modelli strutturali su larga scala hanno avuto prestazioni non buone in termini di accuratezza previsionale al di fuori del campione. Infatti, anche McNees (1986) ha dimostrato come per alcune variabili le previsioni siano riprodotte con maggiore accuratezza attraverso l'utilizzo dei modelli VAR rispetto a molte altre specificazioni strutturali.

Oltre a segnalare i pregi dell'applicazione dei processi VAR rispetto ad altre classi di modelli adottabili, si ritiene equo sottolinearne anche i limiti.

Un primo svantaggio, che si può evidenziare nell'utilizzo dei modelli autoregressivi vettoriali VAR, è il fatto che questi risultano "a-teorici" in quanto utilizzano poche informazioni teoriche relative alle relazioni tra le variabili per indirizzare la specificazione del modello. D'altra parte, delle valide restrizioni di esclusione che assicurino l'identificazione delle equazioni di un sistema strutturale simultaneo informerebbero sulla struttura del modello. Una conseguenza di tale circostanza è che i modelli VAR risultino essere meno inclini all'analisi teorica e perciò alle prescrizioni politiche. Risulta esserci anche un'alta probabilità per cui attraverso l'applicazione dei processi VAR si potrebbe ottenere essenzialmente una relazione spuria andando ad estrarre i dati. In ultimo, spesso non risulta essere chiaro come i coefficienti stimati del processo VAR debbano essere interpretati.

Un secondo svantaggio, che è evidenziabile nell'applicazione dei modelli vettoriali autoregressivi, risulta essere la quantità di parametri da stimare. Infatti, se assumiamo avere *n* equazioni, in cui in ognuna abbiamo *n* variabili e consideriamo *p* ritardi di ciascuna variabile in ogni equazione, allora dovranno essere stimati ($n + pn^2$) parametri. Quindi, per campioni con dimensioni relativamente piccole, i gradi di libertà si esauriranno rapidamente, implicando errori standard abbastanza rilevanti e perciò ampi intervalli di confidenza per i parametri del modello.

Infine, un'ulteriore limitazione che può essere riscontrata nell'utilizzo dei processi VAR risulta essere la condizione necessaria di stazionarietà delle componenti inserite nel modello. Infatti, se si desidera utilizzare test d'ipotesi, singolarmente o

congiuntamente, per esaminare la significatività statistica dei coefficienti, allora è essenziale che tutte le componenti del VAR siano stazionarie. Le componenti da introdurre in un processo VAR(*p*) possono essere trasformate applicando la differenza logaritmica per ricondursi ad una serie storica stazionaria. Tuttavia, i promotori dell'approccio VAR sottolineano che l'applicazione della differenza logaritmica, per indurre una serie storica ad essere stazionaria, non dovrebbe essere apportata. Essi sostengono infatti che lo scopo della stima del VAR è puramente quello di esaminare le relazioni tra le variabili, e che la differenziazione possa sottrarre informazioni su qualsiasi relazione a lungo termine tra le serie storiche.

1.2 TECNICHE DI MACHINE LEARNING NON PARAMETRICHE

Negli ultimi anni si è registrato un grande aumento dell'uso del Machine Learning, tale aumento è dovuto alla sua versatilità che consente applicazioni in molteplici ambiti. Riportando una delle definizioni più ricorrenti e diffuse in letteratura, il Machine Learning, secondo Tom M. Mitchell[9], è: "un programma informatico che apprende dall'esperienza *E* rispetto ad alcune classi di compiti *T* e alla misura delle prestazioni *P*, se la prestazione al compito *T*, misurato da *P*, migliora con l'esperienza *E*." Infatti, il Machine Learning studia come la macchina può apprendere (o migliorare le proprie prestazioni) basandosi interamente sui dati a

[9] *Tom M. Mitchell, "Machine Learning", pp.2.*

disposizione, con l'obiettivo di riconoscere modelli complessi e prendere decisioni intelligenti.

In questa tesi, ci concentreremo in particolare su tecniche di Machine Learning relative alla classificazione di dati.

Concettualmente, la classificazione è una forma di analisi dei dati che estrae modelli tentando di descrivere importanti classi di strutture di dati. Tali modelli, denominati classificatori, predicono le "classi etichetta" di variabili categoriche (discrete, non ordinate) migliorando la comprensione dei dati attraverso specifiche analisi. Molti metodi di classificazione sono stati proposti dai ricercatori nell'apprendimento automatico, nel riconoscimento dei modelli e in statistica. In particolare, recenti ricerche di data mining si sono basate su algoritmi residenti in memoria, sviluppando tecniche di classificazione e previsione scalabili in grado di gestire grandi quantità di dati.

Tuttavia, non è possibile delineare un preciso ambito di applicazione dei metodi di classificazione, ne esistono, infatti, molteplici, tra cui: il rilevamento delle frodi, il marketing mirato, la previsione delle prestazioni, la produzione e la diagnosi medica.

In generale, la classificazione dei dati è un processo che consiste in due differenti fasi: la fase di apprendimento, dove avviene la costruzione di un modello di classificazione, e la fase di classificazione, dove il modello viene utilizzato per prevedere le "classi etichetta" cui si riferiscono i dati a disposizione.

Nella prima fase, viene costruito un classificatore che va a descrivere un predeterminato set di classi di dati o concetti. Questa è la fase di apprendimento (o di "allenamento") dove un algoritmo di classificazione viene utilizzato per costruire un classificatore, analizzando o apprendendo da un set di apprendimento costituito da database di strutture di dati e dalle "classi etichetta" associate.

Una struttura di dati è rappresentata da un vettore di attributi *n*-dimensionale, $X = (x_1, \dots, x_n)$ raffigurante *n* misurazioni effettuate sulla struttura dei dati derivanti da *n* database di attributi, $A_1, \dots, A_n$. Ogni struttura di dati si presume appartenga ad una classe predefinita come determinato da un altro attributo del database denominato l'attributo della "classe etichetta". L'attributo della "classe etichetta" risulta essere discreto, ordinato, ed è categorico (o nominale) nel senso che ogni valore viene utilizzato come categoria o classe. Le singole strutture di dati che costituiscono il set di apprendimento sono denominate strutture di dati di apprendimento[10] e vengono campionate casualmente dal database analizzato.

La prima fase del processo di classificazione può anche essere intesa come una mappatura o l'individuazione di una funzione, $y = f(X)$, la quale può predire la "classe etichetta" di una data struttura di dati. In questa fase, la volontà è quella di "apprendere" tale funzione, tipicamente rappresentata sotto forma di regole di

[10] Nella letteratura relativa al Machine Learning, le strutture di dati di apprendimento si riferiscono comunemente a campioni di apprendimento.

classificazione, alberi di decisione o formule matematiche, che separa i dati in classi.

Tuttavia, a questo punto è d'obbligo sottolineare la distinzione del caso in cui la "classe etichetta" di ogni struttura di dati d'apprendimento viene fornita, dal caso in cui non si è a conoscenza di quest'ultima e il numero o l'insieme di classi che debbono essere apprese potrebbero non essere noti in anticipo.

Nel primo caso si fa riferimento all'*apprendimento supervisionato*, il quale è fondamentalmente sinonimo di classificazione. La supervisione nell'apprendimento avviene attraverso la conoscenza anticipata delle "classi etichetta" nel set di dati d'apprendimento.

Il secondo caso, invece, è relativo all'*apprendimento non supervisionato*, il quale è facilmente associabile alla clusterizzazione. In contrasto con l'apprendimento supervisionato, l'applicazione dell'apprendimento non supervisionato è spesso molto più laboriosa in quanto l'analisi risulta essere maggiormente soggettiva non avendo un obiettivo d'analisi o una risposta previsionale ben definita. Inoltre, può essere molto complesso valutare i risultati ottenuti mediante apprendimento non supervisionato, in quanto, non esiste un meccanismo universalmente riconosciuto per effettuare una convalida incrociata o validare i risultati su un set di dati indipendenti.

Infine, esiste un caso intermedio denominato *apprendimento semi-supervisionato* che comprende una classe di tecniche di Machine Learning che utilizza sia strutture

di dati etichettate che strutture di dati non etichettate per l'apprendimento di un modello. In questo approccio, le strutture di dati etichettate vengono utilizzate per apprendere i modelli delle classi mentre le strutture di dati non etichettate vengono adottate per raffinare i confini tra le classi determinate.

Nela seconda fase del processo di classificazione, il modello appreso viene utilizzato per classificare i dati d'interesse. In primo luogo, viene stimata l'accuratezza predittiva del classificatore. Se quest'ultima dovesse essere stimata attraverso l'utilizzo del set di apprendimento, tale stima sarebbe probabilmente molto ottimista, poiché il classificatore tende a sovrabbondare con i dati. Per questo motivo, viene utilizzato un set di dati e le loro "classi etichetta" associate per testare il classificato, denominato struttura di dati test. Quest'ultime risultano essere indipendenti dalle strutture di dati d'apprendimento, ciò comporta il fatto che non vengono utilizzati per la costruzione del classificatore.

1.2.1 K-Nearest-Neighbor

A proposito della trattazione generale della tecnica di classificazione e delle fasi del processo che la compongono, l'attenzione viene rivolta verso una tecnica di classificazione in particolare: il K-Nearest-Neighbor.

Il classificatore K-Nearest-Neighbor è una tecnica di "*lazy-learner*"[11], in cui il metodo di apprendimento attende fino all'ultimo istante prima di effettuare la costruzione di un modello per classificare la struttura di dati test. Infatti, quando viene analizzata una struttura di dati di apprendimento, un "*lazy-learner*" semplicemente memorizza quest'ultima (o processa solo in minima parte) e attende fino all'istante in cui viene data da analizzare una struttura di dati test. Quindi, solo quando il classificatore si trova ad analizzare una struttura di dati test questo esegue una generalizzazione per classificare la struttura di dati test basandosi sulle similarità della struttura con dati di allenamento immagazzinati in precedenza.

Diversamente, i metodi "*eager-learner*[12]", quando ricevono una struttura di dati d'apprendimento, costruiranno un modello di generalizzazione prima di ricevere nuove strutture di dati da classificare.

A differenza di quest'ultimo approccio, il metodo "*lazy-learner*" effettua un'analisi meno laboriosa quando viene proposta la struttura di dati di apprendimento mentre il processo risulta più impegnativo quando si effettua la classificazione o la previsione numerica.

Il metodo K-Nearest-Neighbor (KNN) è stato descritto per la prima volta agli inizi del 1950, tuttavia, essendo questo un metodo molto laborioso quando analizza un

[11] *Han, Kamber, Pei, "Data Mining: Concept and Techniques", Third Edition, pp. 423.*
[12] *Han, Kamber, Pei, "Data Mining: Concept and Techniques", Third Edition, pp. 423.*

vasto set di dati, esso non ha ottenuto popolarità fino agli inizi del 1960 quando l'evoluzione tecnologica ha reso possibile ridurre i tempi di calcolo.

Il classificatore KNN si basa su un apprendimento per analogia, che consiste nella comparazione di una data struttura di dati test con delle strutture di dati d'apprendimento risultanti simili. Ogni struttura di dati, come già discusso in precedenza, è descritta da *n* attributi e ogni struttura di dati rappresenta un punto in uno spazio *n*-dimensionale. In questo modo, tutte le strutture di dati d'apprendimento sono memorizzate in uno spazio del modello. Quando si considera una struttura di dati sconosciuta, un classificatore K-Nearest-Neighbor va alla ricerca dello spazio del modello per le *k* strutture dei dati che sono le più vicine alla struttura di dati sconosciuta. Queste *k* strutture di dati sono i *k* "*nearest neighbors*" della struttura di dati sconosciuta.

La vicinanza, in questo contesto s'intende in termini di una metrica, come ad esempio la distanza Euclidea. Dati due punti o due strutture di dati, $X_1 = (x_{11}, \dots, x_{1n})$ e $X_2 = (x_{21}, \dots, x_{2n})$, la distanza Euclidea tra X_1 *e* X_2 risulta essere:

$$Distanza\ Euclidea(X_1, X_2) = \sqrt{\sum_{i=1}^{n} (x_{1i} - x_{2i})^2} \qquad (1.26)$$

Tuttavia, in questo contesto si può considerare l'applicazione di differenti distanze come, ad esempio, la similarità del coseno, la distanza di Minkowsky e la distanza di correlazione, le quali vengono illustrate in seguito:

$$Similarità\ del\ coseno(X_1, X_2) = \frac{X_1 \cdot X_2}{|X_1|\ |X_2|} \quad (1.27)$$

$$Distanza\ di\ Minkowsky(X_1, X_2) = \left(\sum_{i=1}^{n} |x_{1i} - x_{2i}|^r\right)^{\frac{1}{r}} \quad (1.28)$$

$$Distanza\ di\ correlazione(X_1, X_2) = \frac{\sum_{i=1}^{n}(x_{1i} - \mu_{1i})(x_{2i} - \mu_{2i})}{\sqrt{\sum_{i=1}^{n}(x_{1i} - \mu_{1i})^2(x_{2i} - \mu_{2i})^2}} \quad (1.29)$$

dove μ_{1i} e μ_{2i} risultano essere i valori medi rispettivamente della struttura dati $X_1\ e\ X_2$.

Tipicamente, prima di calcolare le distanze tra le strutture di dati d'apprendimento e la struttura di dati test, è necessario effettuare una trasformazione dei dati a disposizione applicando una normalizzazione.

La trasformazione dei dati mediante normalizzazione ha l'obiettivo di evitare che la dipendenza della scelta dell'unità di misura infici i risultati dell'analisi, tentando di dare ad ogni attributo lo stesso peso. Proprio per questo motivo, l'applicazione delle tecniche di normalizzazione è molto diffusa all'interno degli algoritmi di classificazione tra cui: le reti neurali e le tecniche di apprendimento, le quali si basano sulla misurazione di distanze come il KNN e la clusterizzazione.

Tra le molteplici tecniche di normalizzazione può essere citata la *normalizzazione min-max* che effettua una trasformazione lineare dei dati originali.

Supponendo che min_A e max_A siano il minimo e il massimo valore dell'attributo A, la normalizzazione min-max trasforma un valore v_i dell'attributo A in v_i' con la seguente formula:

$$v_i' = \frac{v_i - min_A}{max_A - min_A} \tag{1.30}$$

La normalizzazione min-max, illustrata in precedenza, tenderà quindi a preservare le relazioni appartenenti ai dati originali.

La seconda tecnica di trasformazione dei dati, che verrà presentata, fa riferimento alla *normalizzazione z-score* (o normalizzazione a media zero), in cui i valori per un attributo A, vengono normalizzati basandosi sulla media e la deviazione standard dello stesso A. Un valore, v_i dell'attributo A, viene normalizzato in v_i' mediante la seguente formula:

$$v_i' = \frac{v_i - \bar{A}}{\sigma_A}, \tag{1.31}$$

dove, $\bar{A}$ e σ_A sono rispettivamente la media e la deviazione standard dell'attributo A. Questa tecnica di normalizzazione risulta molto utile nel momento in cui non si è a conoscenza del valore massimo e minimo dell'attributo A, o quando si è in presenza di valori anomali (outliers) che dominano nella normalizzazione min-max. Tuttavia, esiste anche una variazione di quest'ultima tecnica che va a rimpiazzare la deviazione standard inserendo la deviazione media assoluta trasformando la precedente formula di normalizzazione come segue:

$$v_i' = \frac{v_i - \bar{A}}{s_A}, \tag{1.32}$$

dove $s_A = \frac{1}{n}(|v_1 - \bar{A}| + \cdots + |v_n - \bar{A}|)$.

La deviazione media assoluta risulta essere più robusta ai valori anomali rispetto alla deviazione standard classica, poiché alle deviazioni dalla media non viene applicato il quadrato, riducendo quindi gli effetti anomali.

La terza e ultima tecnica di normalizzazione, che verrà presentata, si riferisce alla *normalizzazione mediante scala decimale*, la quale prevede lo spostamento del punto decimale nei valori dell'attributo A. In questo caso, il numero degli spostamenti del punto decimale dipende dal massimo dell'attributo A in valore assoluto. Infatti, un valore v_i dell'attributo A, viene normalizzato a v_i' attraverso il seguente calcolo:

$$v_i' = \frac{v_i}{10^j}, \tag{1.33}$$

dove *j* è il più piccolo numero intero tale che $\max(|v_i'|) < 1$.

I classificatori KNN possono essere utilizzati anche per effettuare delle previsioni numeriche, cioè per restituire una previsione di valori reali per una data struttura di dati non conosciuta. In questo caso, il classificatore restituisce il valore medio delle etichette dei valori reali associati alle *k* strutture di dati più simili alla struttura di dati sconosciuta.

Emerge, a questo punto, il problema di come determinare il valore *k*. Tuttavia, la letteratura non fornisce una regola precisa ma, indica che questo valore inserito

nella tecnica di classificazione KNN debba essere individuato empiricamente. Ciò sta a significare che, partendo da $k = 1$ si utilizza il set di dati test per misurare la percentuale d'errore commessa dal classificatore. Questo processo può essere ripetuto incrementando ogni volta il valore di *k* e calcolando l'errore commesso dal classificatore, andando infine a selezionare quel valore di k che rende minimo l'errore.

Capitolo 2

ANALISI DEI DATI E METODOLOGIE UTILIZZATE

2.1 MERCATO DELLE CRIPTOVALUTE

Il mercato delle criptovalute continua ad attirare l'attenzione di investitori, imprenditori, autorità di regolamentazione e degli operatori finanziari in generale. Recenti discussioni pubbliche riguardanti le criptovalute sono state innescate dai sostanziali cambiamenti dei loro prezzi, sostenendo che il mercato delle criptovalute risulta essere una bolla senza alcun valore fondamentale e facendo sorgere dubbi in relazione alla loro evasione dalla normativa e dalla supervisione legale.

Le criptovalute sono attività finanziarie digitali, per le quali le registrazioni e i trasferimenti di proprietà vengono garantiti da una tecnologia crittografica piuttosto che da una banca o una terza parte.

Infatti, alla fine del 2008 l'inventore dei Bitcoin, noto con lo pseudonimo di Satoshi Nakamoto, introdusse un nuovo sistema crittografico decentralizzato, alla base della tecnologia blockchain[13]. Contemporaneamente, lo stesso autore propose l'uso della tecnologia blockchain per la gestione della criptovaluta comunemente

[13] *Satoshi Nakamoto, "Bitcoin: A Peer-to-Peer Electronic Cash System"*, www.bitcoin.org.

conosciuta con il nome di Bitcoin. Il lavoro di Nakamoto è considerato rivoluzionario in quanto ha risolto alcune questioni come il problema della doppia spesa, il problema degli attacchi hacker causati della centralità della rete e i lunghi periodi associati a transazioni transfrontaliere e interbancarie. L'innovazione chiave consiste nell'utilizzare un sistema di calcolo distribuito (denominato algoritmo "Proof-of-Work") per condurre una "elezione" globale ogni 10 minuti, consentendo alla rete decentralizzata di arrivare ad un "consenso" sullo stato delle transazioni. Né Satoshi Nakamoto, né nessun altro esercita il controllo individuale sul sistema Bitcoin, il quale opera sulla base di principi matematici completamente trasparenti, codici open source e consenso tra i partecipanti.

Secondo Antonopoulos[14] "Bitcoin è una raccolta di concetti e tecnologie che formano le basi per un ecosistema di denaro digitale." Infatti, le unità di valuta chiamate Bitcoin vengono utilizzate per immagazzinare e trasferire valore tra i partecipanti del network Bitcoin. Gli utenti Bitcoin comunicano tra loro utilizzando il protocollo Bitcoin principalmente attraverso internet, sebbene possano essere utilizzate altre reti di trasporto.

Negli anni successivi alla nascita di Bitcoin, molte altre criptovalute (denominate *altcoins*), come Ethereum[15] e Litecoin[16], sono state sviluppate.

[14] *Antonopoulos, "Mastering Bitcoin: Programming the Open Blockchain", O'Reilly.*
[15] https://www.ethereum.org/.
[16] https://www.litecoin.org/.

Spesso, queste criptovalute sono state sviluppate per scopi diversi o cercando di migliorare alcune limitazioni di Bitcoin, come ad esempio la fornitura limitata di Bitcoin, l'elevato consumo energetico della rete o il meccanismo "Proof-of-Work" per approvare il consenso degli utenti.

Riferendosi alle criptovalute, si è spesso discusso relativamente la loro collocazione tra le classi di attività. Infatti, anche se queste vengono considerate valute, nel senso che sono mezzi digitali di scambio, sorgono alcuni limiti a riguardo. Se si considerano le valute tradizionali, il motivo principale che spinge gli individui a utilizzare una valuta consolidata come il dollaro statunitense (USD) o l'euro (EUR) è che il suo valore rimane relativamente coerente nel tempo e che un governo agisce come garante.

Le criptovalute sono carenti in entrambi gli elementi sopra considerati. Questo fa sì che il mercato delle criptovalute sia estremamente volatile e attualmente rende le criptovalute inadatte come memorizzazione affidabile del valore o come mezzo di scambio (Ciaian,2016). Secondo Yermack (2015), la scarsità di Bitcoin e la sua instabilità intrinseca sono le principali ragioni per cui non può essere classificata come una valuta reale, applicando tale considerazione anche a molte altre criptovalute. Inoltre, i dati di Chainanalysis nel 2018 indicano che la maggior parte degli investitori non utilizza Bitcoin come mezzo di scambio, ma piuttosto lo vedono come uno strumento d'investimento. Infatti, lo stesso studio indica che 6 milioni di Bitcoin sono detenuti da investitori a lungo termine (>1 anno) rispetto a

circa 5 milioni di Bitcoin che sono detenuti invece da speculatori a breve termine (<1 anno), con i restanti 10 milioni che non sono ancora stati estratti o vengono considerati persi. I dati indicano inoltre che la stragrande maggioranza delle transazioni di Bitcoin avvengono nei mercati finanziari e che Bitcoin viene raramente utilizzato per pagare beni o servizi (Murphy, 2018). Tuttavia, tali risultati, ottenuti da specifici studi su Bitcoin, sono di dubbia generalizzazione a tutte le criptovalute esistenti soprattutto perché ognuna ha uno scopo specifico e caratteristiche distinte come l'offerta, la domanda e il volume delle transazioni.

Quindi, le criptovalute possono essere considerate attività finanziarie perché hanno un valore per i detentori di criptovalute, anche se non rappresentano una corrispondente passività e non sono garantite da alcuna attività fisica di valore (come ad esempio l'oro o lo stock di attrezzature di un'impresa).

L'utilizzo delle criptovalute come strumento di copertura contro l'incertezza politica e finanziaria del mercato è un argomento largamente studiato. Infatti, vari studi, come Brière (2015), Dyrberg (2016), Li e Wang (2017) e Bouri (2017) hanno scoperto che Bitcoin può essere utilizzato efficacemente come strumento di copertura contro l'incertezza globale e che esso rappresenta un buon elemento di diversificazione all'interno di un portafoglio d'investimento capace di diversificare una vasta gamma di indici, valute e materie prime.

Di seguito vengono riportate le 10 criptovalute maggiormente capitalizzate con l'aggiunta di ZCash che sarà oggetto di studio nei successivi paragrafi ma che risulta essere collocata al quarantesimo posto per capitalizzazione di mercato.

Tabella 2.1 - Migliori criptovalute per capitalizzazione di mercato al 24 dicembre 2020. (Fonte: https://coinmarketcap.com).

#	*Nome*	*Capitalizzazione di Mercato*	*Prezzo*	*Volume (24h)*	*Offerta in Circolazione*
1	**Bitcoin***	$447,913,320,650	$23,735.95	$42,438,466,727	18,580,781 BTC
2	**Ethereum***	$71,489,909,178	$611.61	$13,467,430,029	113,983,877 ETH
3	**Tether**	$20,683,694,686	$1.00	$68,670,449,681	20,684,166,934 USDT
4	**Xrp**	$14,323,290,722	$0.337819	$15,968,693,956	45,404,028,640 XRP
5	**Litecoin***	$8,386,711,221	$111.57	$11,726,335,900	66,169,961 LTC
6	**Bicoin Cash**	$5,956,956,931	$296.28	$5,546,226,793	18,600,863 BCH
7	**Cardano**	$4,872,799,246	$0.152883	$1,103,134,758	31,112,484,646 ADA
8	**Binance Coin**	$4,788,563,375	$32.50	$420,094,324	144,406,561 BNB
9	**Polkadot**	$4,641,501,504	$5.13	$506,425,344	893,681,123 DOT
10	**Chainlink**	$4,620,664,595	$11.58	$1,384,513,016	398,509,556 LINK
40	**ZCash***	$673,087,240	$60.85	$391,589,214	10,768,388 ZEC

Nella Tabella 2.1 ci si riferisce alla capitalizzazione di mercato come al valore totale delle criptovalute al prezzo di mercato. Invece, il volume è il totale delle transazioni nelle ultime 24 ore e l'offerta in circolazione è l'ammontare nominale delle criptovalute disponibili. È facilmente riscontrabile come il mercato delle criptovalute sia maggiormente rappresentato da Bitcoin, infatti la capitalizzazione di mercato di quest'ultimo risulta essere sei volte la capitalizzazione di mercato della seconda criptovaluta più capitalizzata, Ethereum.

Questa notevole differenza tra le criptovalute esistenti è riscontrabile anche nel prezzo di tali *asset* e nella quantità di transazioni effettuate.

In questa sede verrà considerato il mercato delle criptovalute facendo riferimento solamente ad alcune delle valute digitali che realmente vengono scambiate.

Di seguito è rappresentato l'andamento dei prezzi di chiusura giornalieri delle criptovalute analizzate in questa tesi: Bitcoin, Ethereum, Litecoin, ZCash.

Figura 2.1 - Prezzi di chiusura delle criptovalute analizzate dal 29 ottobre 2016 al 26 dicembre 2020. (Fonte: https://it.finance.yahoo.com).

Risulta molto facile percepire l'alta volatilità delle criptovalute prese in esame notando in particolare in Bitcoin il repentino susseguirsi di trend crescenti e trend

decrescenti. Le altre tre criptovalute risultano essere prezzate dai mercati di scambio in maniera molto inferiore rispetto alla principale criptovaluta. Tuttavia, può essere sottolineata una particolarità nella criptovaluta ZCash che nel primo giorno della sua emissione ha raggiunto il suo massimo valore ($2044,47), dovuto probabilmente all'iniziale entusiasmo dei mercati per l'asset, per poi scendere vertiginosamente durante tutta la sua esistenza.

Relativamente ai volumi delle criptovalute scambiate, in seguito viene rappresentato il grafico relativo al periodo 29 ottobre 2016 - 26 dicembre 2020.

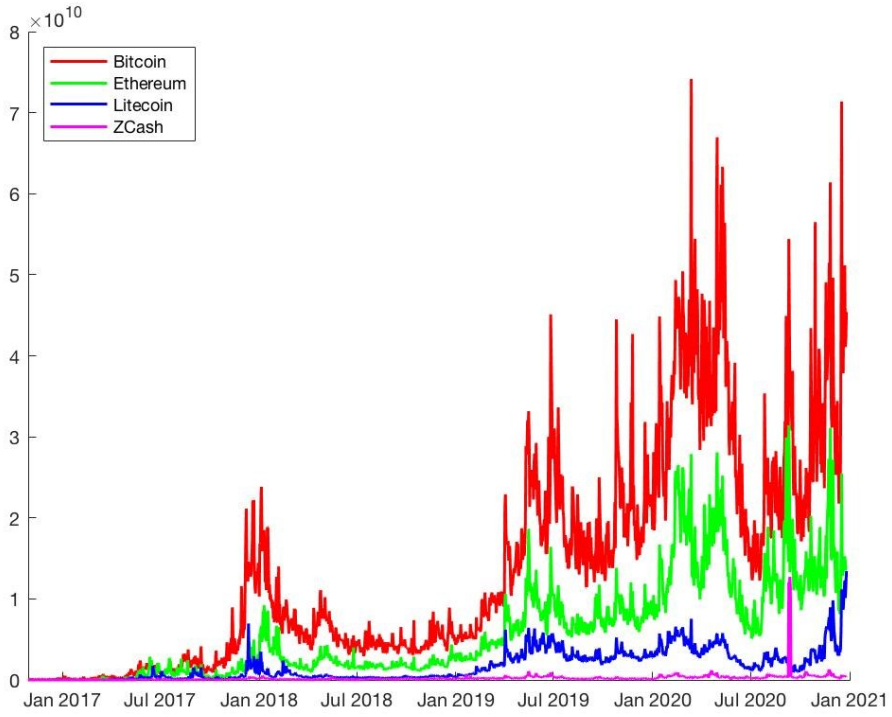

Figura 2.2 - Volumi delle criptovalute analizzate dal 29 ottobre 2016 al 26 dicembre 2020. (Fonte: https://it.finance.yahoo.com).

Anche in questo caso è indubbia la differenza tra la quantità di transazioni concernenti Bitcoin e le transazioni rispettivamente di Ethereum, Litecoin e ZCash. Una particolarità sottolineabile persino nei volumi delle criptovalute ricade su ZCash, che nel periodo tra Luglio 2020 e Dicembre 2020 ha raggiunto il proprio valore massimo riuscendo a superare le quantità scambiate di Litecoin e avvicinarsi a quelle di Ethereum, seppur come abbiamo visto in precedenza, la capitalizzazione di mercato di ZCash risulti essere notevolmente inferiore.

2.1.1 Collezione di dati finanziari sulle criptovalute

La scelta di considerare solamente alcune criptovalute è dovuta al fatto che la disponibilità dei dati finanziari ad alta frequenza utilizzati nell'analisi e nella previsione in questo elaborato risulta essere relativamente scarsa per alcune criptovalute. Perciò, si è deciso di prendere in considerazione quattro principali criptovalute che si contraddistinguono ognuna per specifiche caratteristiche: Bitcoin, Ethereum, Litecoin, ZCash.

Oltre a Bitcoin (BTC), la quale è stata ampiamente descritta nel paragrafo precedente, Ethereum (ETH) è la seconda criptovaluta per capitalizzazione di mercato e fama, la quale venne lanciata nel 2015. Quest'ultima è una piattaforma software open source, basata sulla blockchain, utilizzata per la propria criptovaluta, Ether. Ethereum consente di costruire ed eseguire *smartcontracts* e *distributed applications* senza alcuna interruzione, frode, controllo o interferenza da terze parti.

La terza criptovaluta per capitalizzazione di mercato considerata è Litecoin (LTC). Quest'ultima venne lanciata nel 2011 in seguito ad una fork dalla blockchain di Bitcoin da cui possiamo evidenziare delle differenze principalmente nel tempo di creazione dei blocchi, che risulta essere inferiore in Litecoin, consentendo una più veloce conferma delle transazioni. Tuttavia, l'attività di mining della criptovaluta risulta essere più difficile e costosa rispetto al protocollo riferito a Bitcoin.

La quarta e ultima criptovaluta analizzata in questa sede è ZCash. Questa venne emessa per la prima volta alla fine del 2016 e fu costruita sul codice blockchain di Bitcoin, ma si distingue da quest'ultimo per la sua elevata attenzione alla privacy e sicurezza. La principale accortezza della piattaforma sottostante alla criptovaluta in questione è la garanzia della privacy. Infatti, ZCash fa uso di un meccanismo "Zero-Knowledge-Proof" per garantire la validità delle transazioni mantenendo la riservatezza sull'importo e sull'emittente. I dati nel presente elaborato sono stati scaricati dalla piattaforma CryptoDataDownload[17] prendendo in considerazione il mercato Gemini. Sono stati collezionati i dati relativamente alle quattro criptovalute descritte in precedenza considerando come riferimento temporale il minuto. Infatti, sono stati memorizzati Prezzo Massimo, Prezzo Minimo, Prezzo di Chiusura e Volume per ogni minuto nel periodo temporale che va dal 27 settembre 2020 al 3 ottobre 2020.

[17] https://www.cryptodatadownload.com/index.html

Questa scelta è strettamente dipendente dalle caratteristiche intrinseche dei dati ottenibili attraverso Twitter che verranno discussi nei successivi paragrafi.

Di seguito vengono mostrati i grafici dei prezzi di chiusura delle quattro criptovalute considerate ad alta frequenza per l'intervallo temporale 27 settembre 2020 – 3 ottobre 2020.

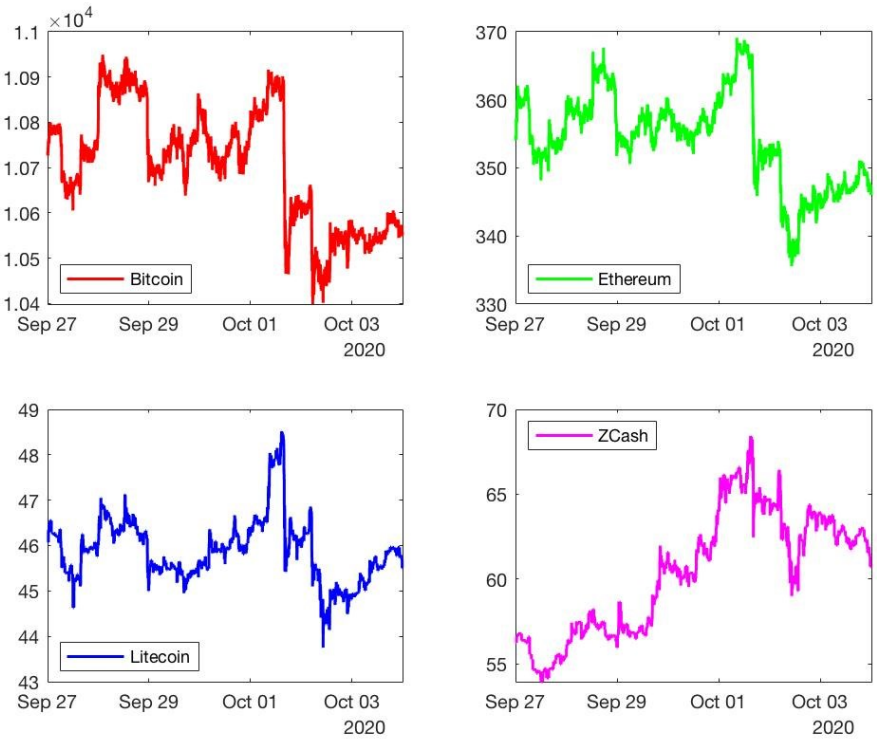

Figura 2.3 - Prezzi al minuto delle quattro criptovalute analizzate nel periodo 27 settembre 2020 – 3 ottobre 2020. (Fonte: https://www.cryptodatadownload.com/index.html).

Anche in questo caso, come nel precedente grafico dei prezzi di chiusura giornalieri, è evidenziabile un'alta volatilità degli strumenti riscontrata dalle ampie oscillazioni anche all'interno delle singole giornate di scambio.

Inoltre, è possibile notare un comportamento similare tra le due criptovalute maggiormente capitalizzate: Bitcoin ed Ethereum. Oltre ai prezzi di riferimento delle criptovalute sono stati presi in considerazione nelle analisi e nelle previsioni effettuate i dati dei volumi essendo questi riferiti alle transazioni effettuate e quindi di notevole importanza. Di seguito verranno mostrati i grafici relativi ai volumi delle quattro criptovalute analizzate nel periodo 27 settembre 2020 – 3 ottobre 2020 con frequenza 1 minuto per un totale di 10.080 dati.

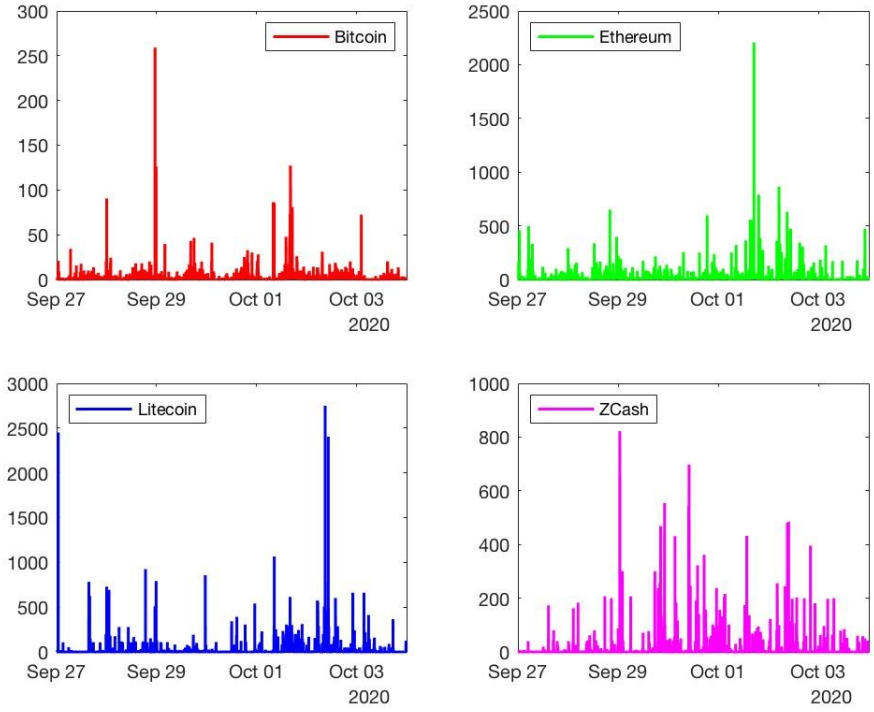

Figura 2.4 - Volumi al minuto delle quattro criptovalute analizzate nel periodo 27 settembre 2020 - 3 ottobre 2020. (Fonte: https://www.cryptodatadownload.com/index.html).

In questo caso, trattandosi dei volumi delle criptovalute considerate, è possibile notare come transazioni di diverso ammontare si susseguono sottolineando ancora una volta l'alta volatilità e la velocità di scambio anche in lassi temporali relativamente ridotti come il minuto.

2.2 SENTIMENT ANALYSIS DEI TWEET

La crescente popolarità di Bitcoin e delle criptovalute in generale ha generato interesse nell'individuare i fattori trainanti nella formazione del loro prezzo. In letteratura i principali elementi identificati come determinanti nella formazione del prezzo delle criptovalute sono: la forza della domanda e dell'offerta nel mercato delle criptovalute, la loro attrattività e lo sviluppo macro-finanziario globale.

Infatti, lo studio di Buchholz (2012) evidenzia come un'importante determinante del prezzo di Bitcoin l'interazione tra domanda e offerta, dove quest'ultima determina l'ammontare di unità in circolazione e quindi la scarsità nel mercato.

Secondo Kristoufek (2013), la formazione del prezzo di Bitcoin non può essere spiegata da teorie economiche standard come il modello dei flussi di cassa futuri, la parità del potere d'acquisto, la parità scoperta dei tassi d'interesse, perché molte caratteristiche della domanda e dell'offerta possedute dalle valute tradizionali, non sono riscontrabili nel mercato delle criptovalute.

Invece, Van Wijk (2013) sottolinea come lo sviluppo macro-finanziario globale, rappresentato, ad esempio, da indici di borsa, tassi di cambio e prezzo del petrolio,

giochi un ruolo fondamentale nella determinazione del prezzo di Bitcoin. Lo stesso studio evidenzia che, in particolare, l'indice Dow Jones, il tasso di cambio euro-dollaro e il prezzo del petrolio hanno un impatto significativo sul valore di Bitcoin nel lungo periodo.

Tuttavia, la prevedibilità del mercato delle criptovalute dovrebbe essere notevole, in quanto, secondo le Ipotesi dei Mercati Efficienti (EMH) un mercato prevedibile è inefficiente in modo informativo in quanto le informazioni disponibili non vengono pienamente riflesse dai prezzi di mercato. L'inefficienza del mercato è supportata da alcune anomalie riscontrabili nel mercato di Bitcoin, o in generale nel mercato degli *altcoin*, in cui esiste una forte interdipendenza e correlazione maggiormente intensa nel breve periodo. Infatti, secondo le Ipotesi dei Mercati Efficienti, in un mercato efficiente successivi cambiamenti di prezzo sono indipendenti per definizione (Fama, 1970). Inoltre, un'altra importante assunzione delle Ipotesi dei Mercati Efficienti afferma che gli investitori sono definiti come operatori razionali e il valore degli strumenti finanziari è sancito dal loro valore fondamentale. Tuttavia, un articolo di Silverman (2017) attesta che a causa della mancanza di un valore intrinseco e a causa del prezzo delle criptovalute trainate dalla speculazione, non c'è modo di attribuire un valore fondamentale alle valute digitali, rendendo il mercato irrazionale.

Le Ipotesi dei Mercati Efficienti racchiude la teoria standard neoclassica dei mercati finanziari ma questa si concentra con minore attenzione sul fattore

comportamentale e sugli effetti emotivi che gli operatori del mercato hanno sul mercato.

In questa sede infatti ci interessa analizzare come i nuovi messaggi, le nuove notizie e le emotività degli operatori del mercato in relazione a queste, possono migliorare la previsione e l'analisi dei prezzi e volumi delle criptovalute. Allora, in questo contesto risulta più consono parlare di Ipotesi di Mercato Adattivo (AMH). Il fautore di questa teoria, Lo (2012), sostiene che i mercati non possono sempre essere considerati efficienti, ma competitivi e adattabili, variando il loro grado di efficienza al variare dei comportamenti degli investitori nel tempo. Infatti, secondo le Ipotesi di Mercato Adattivo, l'investitore si comporta in modo razionale solamente nei periodi di certezza, mentre nei periodi di incertezza il suo comportamento risulta di difficile spiegazione in quanto questo è guidato dall'istinto e dall'emozione.

Nell'era di Internet è cambiato il modo in cui gli investitori o gli individui in generale esprimono le loro opinioni o attitudini, utilizzando maggiormente post nei blog, forum online, siti web di recensioni di prodotti e social media.

Proprio social media, come Facebok, Twitter, Google Plus, negli ultimi anni hanno riscontrato particolare successo, in quanto investitori e operatori che agiscono nella finanza esprimono al loro interno emozioni, opinioni e punti di vista circa la loro vita quotidiana.

Risulta quindi a questo punto importante introdurre la Sentiment Analysis (SA), uno strumento capace di trarre informazioni utili e utilizzabili nelle analisi direttamente da testi o messaggi scritti ad esempio nei social media.

La Sentiment Analysis è un processo che automatizza l'estrazione di atteggiamenti, opinioni, punti di vista ed emozioni direttamente da testi, discorsi, tweet e fonti di database attraverso il Natural Language Processing (NLP). La Sentiment Analysis consiste nel classificare le opinioni espresse dal testo in categorie come "positivo", "negativo" o "neutro".

In ambito economico, il sentimento può essere definito come un'errata percezione che può portare alla sbagliata valutazione del valore fondamentale di un determinato asset (Levy, 2010). Il sentimento degli operatori può perciò rendere il prezzo degli strumenti finanziari diverso dal loro valore fondamentale, rispettando un concetto cardine dei mercati finanziari, il quale afferma che il processo decisionale è guidato da fattori psicologici ed emotivi che non sempre rispecchiano il valore fondamentale dell'asset.

Non è possibile prevedere l'arrivo di una nuova notizia così come non è possibile prevedere l'andamento delle criptovalute o in generale degli asset finanziari. Tuttavia, i mercati vengono influenzati dall'arrivo di nuove notizie e quest'ultime costituiscono materiale di studio per la Sentiment Analysis la quale può essere utilizzata per analizzare le emozioni e opinioni degli operatori finanziari.

2.2.1 Collezione dei tweet, pre-processamento e Sentiment Analysis

Con la diffusione dell'utilizzo dei social media, l'informazione circa le opinioni pubbliche è divenuta abbondante. In merito alle criptovalute, social media come Twitter e forum dedicati alle valute digitali risultano essere la fonte primaria di informazione.

In questa sede l'attenzione è stata principalmente rivolta al social media Twitter come piattaforma da cui ottenere dati, da utilizzare nella previsione e nell'analisi delle criptovalute considerate, in quanto Twitter ha ricevuto molta attenzione da parte dei ricercatori negli ultimi anni.

Twitter è un'applicazione di microblogging che permette agli utenti di seguire e commentare il pensiero di altri utenti e condividere con loro opinioni ed emozioni in tempo reale. Infatti, più di un milione di utenti pubblica più di 140 milioni[18] di tweets ogni giorno, rendendo quindi questi dati molto preziosi per i ricercatori che intendono fare previsioni su determinati andamenti.

Relativamente all'abbondanza delle informazioni ottenibili, di seguito viene illustrato un grafico che rappresenta la quantità di tweet pubblicati dagli utenti Twitter circa le quattro criptovalute considerate, Bitcoin, Ethereum, Litecoin, ZCash durante tutto il periodo della loro esistenza.

[18] *Pagolu, Challa, Panda, Majhi, "Sentiment Analysis of Twitter Data for Predicting Stock Market Movements".*

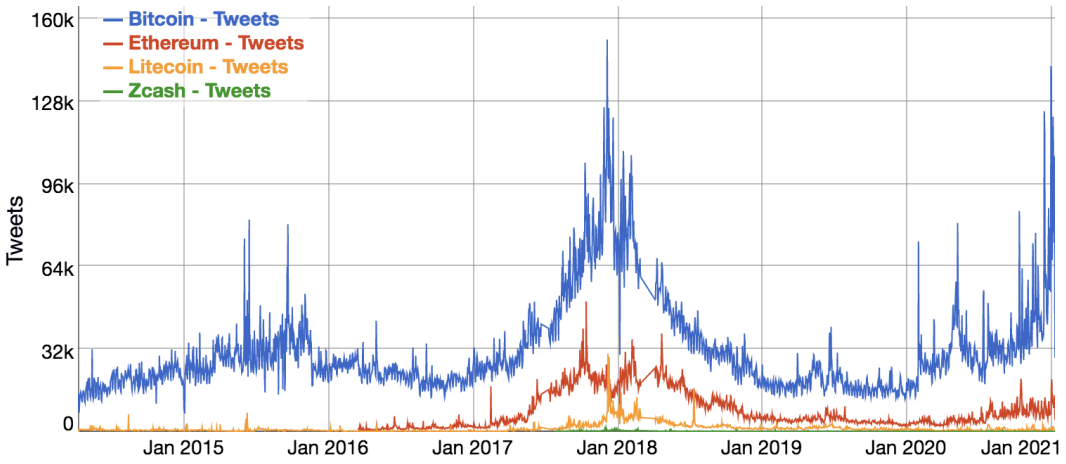

Figura 1.5 - Volume dei tweet per le quattro criptovalute considerate. (Fonte: https://bitinfocharts.com/).

Come è possibile notare, anche per i volumi dei tweet si nota una netta differenza tra le quattro criptovalute analizzate. Infatti, il volume dei tweet di Bitcoin risulta essere molto più elevato, rispetto alle altre criptovalute, potendo sottolineare una proporzione simile alla Figura 2.2, relativa ai volumi delle criptovalute scambiate. Anche in questo caso, alla fine del 2017, si può evidenziare come il volume dei tweet relativi alle quattro criptovalute sia aumentato nel periodo in cui anche i prezzi delle stesse hanno subito degli incrementi notevoli.

A causa delle limitazioni imposte da Twitter nell'ottenimento dei tweet pubblici, in questo studio sono stati collezionati tweet nel periodo temporale 27 settembre 2020 – 3 ottobre 2020. Infatti, Twitter, attraverso il suo metodo di ricerca API, permette

solamente 180 richieste ogni 15 minuti con un limite massimo di 500.000 tweet ottenibili ogni mese per singolo account da sviluppatore[19].

Attraverso il Datafeed Toolbox[20] di Matlab sono stati collezionati tweet in tempo reale al cui interno fosse presente il nome delle criptovaluta presa in considerazione, per un totale di 547.546 tweet ottenuti.

Anche in questo caso, a conferma di quanto evidenziato dalla Figura 2.5, è facilmente riscontrabile la differenza di volume dei tweet ottenuti utilizzando come unica richiesta il nome della criptovaluta: 'bitcoin', 'ethereum', 'litecoin', 'zcash'.

Di seguito la tabella illustra la suddivisione dei tweet per ognuna delle quattro criptovalute considerate nel periodo 27 settembre 2020 – 3 ottobre 2020.

Tabella 2.2 - Numero di tweet per le criptovalute considerate nel periodo 27 settembre 2020 - 3 ottobre 2020.

Criptovaluta	***Tweet***
Bitcoin	421.736
Ethereum	112.024
Litecoin	10.530
ZCash	3.256
Totale	***547.546***

I tweet in genere sono messaggi corti, ciò è dovuto al fatto che Twitter impone come lunghezza massima 140 caratteri per ogni messaggio. Data la natura di questo

[19] https://developer.twitter.com/en
[20] https://it.mathworks.com/products/datafeed.html

servizio di microblogging, gli utenti spesso utilizzano messaggi corti e veloci, acronimi, fanno errori di ortografia, usano emoticon e altri caratteri che esprimono significati speciali.

È comune nel linguaggio di Twitter che gli utenti utilizzino hashtag (#) come prefisso susseguito dall'argomento di discussione oppure il simbolo del dollaro ($) nel momento in cui l'argomento trattato risulti avere carattere finanziario. Oppure, viene spesso utilizzato il simbolo (@) per riferirsi ad altri utenti, avvisandoli automaticamente e aumentando così la visibilità dei tweet emessi.

Proprio per questi motivi e per agevolare la fase di Sentiment Analysis (SA), nel momento in cui sono state effettuate le richieste dei tweet in cui al loro interno vi fossero i nomi delle quattro criptovalute considerate, sono stati specificatamente filtrati i soli tweet emessi in lingua inglese.

Prima di arrivare alla valutazione dei tweet attraverso SA, è inoltre necessaria un'altra fase di notevole rilevanza: il pre-processamento dei dati.

Infatti, i dati di Twitter sono noti per la loro mancanza di struttura e l'alto livello di ridondanza. Di conseguenza, i dati collezionati tramite Twitter richiedono un'ampia elaborazione per renderli utili alla Sentiment Analysis.

Di seguito vengono riportate le fasi di pre-processamento dei tweet svolte per le quattro criptovalute considerate.

Tabella 2.3 - Fasi di pre-processamento dei tweet delle quattro criptovalute considerate.

# *Fase*	*Tecniche di pre-processamento dei tweet*
1	Conversione delle stringhe in minuscolo.
2	Rimozione URLs.
3	Rimozione tag HTML, hashtag (#), ticker ($), menzioni (@) e valori numerici.
4	Eliminazione della punteggiatura.
5	Tokenizzazione di ogni singolo tweet.
6	Rimozione delle parole d'arresto.

Attraverso il Text Analytics Toolbox[21] di Matlab i tweet sono stati pre-processati e nella prima fase, ad esempio, sono stati convertiti tutti i tweet in stringhe con carattere minuscolo. Successivamente sono stati eliminati tutti i componenti URLs, tag HTML, hashtag, ticker, menzioni e valori numerici all'interno dei tweet collezionati. È stata inoltre eliminata la punteggiatura e poi, è stata effettuata la tokenizzazione di ogni tweet, la quale consiste nella suddivisione dei tweet in parole individuali in cui vengono eliminate l'emoticons e altri simboli irrilevanti. Questa fase di pre-processamento è molto importante in quanto permette la successiva rimozione di parole d'arresto che non risultano essere utili alla valutazione del singolo tweet.

A questo punto, è stato possibile effettuare la Sentiment Analysis con l'obiettivo di ottenere un punteggio per ogni tweet relativo alle quattro criptovalute considerate.

[21] https://it.mathworks.com/products/text-analytics.html

Nella letteratura della Sentiment Analysis basata su Twitter, Giachanou e Crestani (2016) hanno individuato quattro approcci per la valutazione dei tweet: l'apprendimento automatico supervisionato, l'approccio basato sul lessico, un approccio ibrido dei due precedenti e un approccio basato sui grafici.

In questa sede è stato utilizzato l'approccio basato sul lessico, in particolare, sempre attraverso il Textanalytics Toolbox di Matlab è stata eseguita la Sentiment Analysis utilizzando il software VADER[22].

Valence Aware Dictionary and sEntiment Reasoner (VADER) è un lessico combinato e un software analitico del sentimento sviluppato da Hutto e Gilbert. VADER è in grado di rilevare la polarità (positiva, neutra, negativa) e l'intensità del sentimento in un determinato testo o documento. Questo potente software analitico è stato sviluppato come soluzione alle difficoltà di analizzare differenti lingue, i simboli e gli stili utilizzati nei testi principalmente nel settore dei social media. I principali obiettivi di VADER illustrati da Hutto e Gilbert sono quelli di far sì che esso funzioni bene sullo stile di testo dei social media e che possa essere prontamente generalizzabile su altri domini. Inoltre, il software, non richiede dati di formazione appresi in precedenza all'utilizzo, risulta essere abbastanza veloce per lo streaming di dati online e non soffre di un trade-off tra velocità e performance.

[22] *Hutto, Gilbert. "VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text".*

Quindi, attraverso VADER, i tweet collezionati sono stati trasformati in degli score appartenenti all'intervallo [-1,1], rappresentando il sentimento relativo a quel tweet nell'istante in cui esso è stato emesso.

2.3 PROCESSI E METODOLOGIE UTILIZZATE

Gli obiettivi principali di questa tesi sono quello di svolgere un'analisi di causalità ed effettuare previsioni con il VAR, migliorando quest'ultime con il metodo K-Nearest Neighbor, utilizzando i dati finanziari relativi alle criptovalute e i tweet associati.

La prima fase di raccolta dei dati relativi alle criptovalute e dei tweet riferiti alle stesse nel periodo 27 settembre 2020 – 3 ottobre 2020 è già stata ampiamente descritta.

Tuttavia, relativamente ai tweet collezionati mediante Twitter API, questi sono attribuiti anche a frazioni di minuto, perciò, per ottenere un unico valore di score riferibile ad un determinato minuto, rendendo i risultati comparabili con quelli dei dati finanziari, è stata effettuata la media degli score relativi ai tweet di un preciso minuto. Invece, per quanto riguarda il volume dei tweet è stato effettuato semplicemente il conteggio dei tweet emessi all'interno di ogni singolo minuto.

Di seguito viene riportata una tabella con alcune statistiche descrittive relative ai tweet ottenuti nel periodo 27 settembre 2020 – 3 ottobre 2020 per le quattro criptovalute considerate.

Tabella 2.4 - Statistiche relative agli score e ai volumi dei tweet al minuto delle criptovalute considerate.

	Numero di tweet	***Media score tweet***	***Deviazione standard score tweet***	***Media volume tweet***	***Deviazione standard volume tweet***
Bitcoin	421.736	0,2083	0,3827	41,84	16,41
Ethereum	112.024	0,2191	0,3890	11,11	6,23
Litecoin	10.530	0,2353	0,3686	1,04	1,21
ZCash	3.256	0,3025	0,3348	0,32	0,76
Media		**0,2413**		**13,58**	

Come si può notare dalla Tabella 2.4, la media degli score attribuiti ai tweet risulta essere sempre positiva in tutte le quattro criptovalute considerate. Infatti, i punteggi sono costantemente distorti in quanto la media delle 4 criptovalute risulta essere pari a 0,2413 all'interno dell'intervallo [-1,1]. Tale risultato è consistente con lo studio di Kennedy e Inkpen (2006), il quale dimostra che gli approcci basati sul lessico, come VADER, tendono ad avere un pregiudizio positivo che può essere attribuito ad una tendenza umana a preferire il linguaggio positivo.

Relativamente alla media dei volumi di tweet emessi per ogni minuto, è osservabile la grande differenza tra le quattro criptovalute mantenendo la proporzione più volte sottolineata in precedenza, che riguarda anche i volumi di valuta digitale scambiata nei mercati.

Di seguito vengono riportati i grafici degli score e dei volumi delle quattro criptovalute considerate nel periodo 27 settembre 2020 – 3 ottobre 2020.

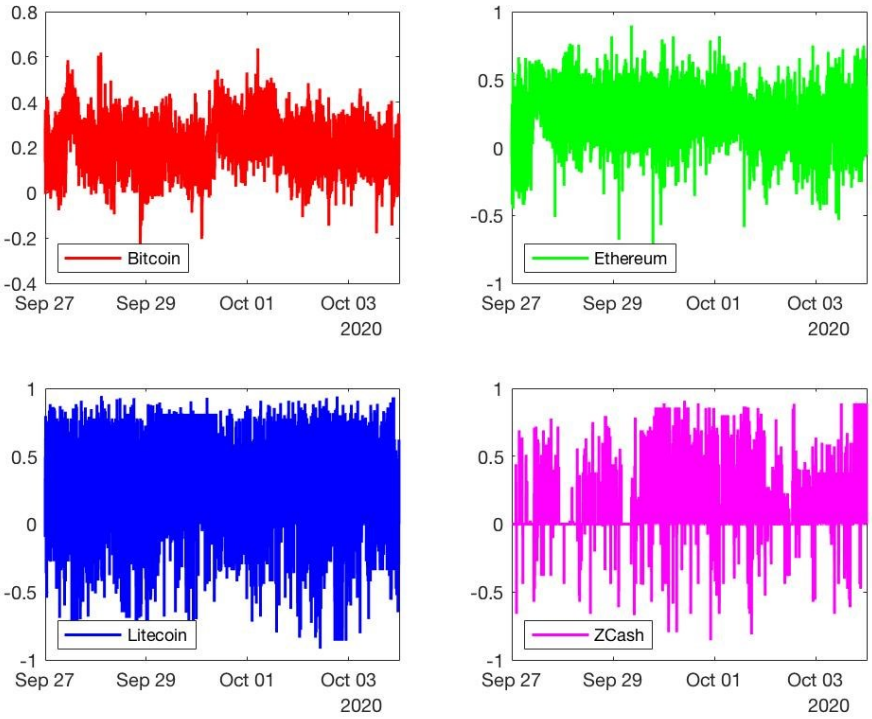

Figura 2.6 - Media degli score al minuto per le quattro criptovalute considerate.

I grafici sopra rappresentati, mostrano gli score relativi ai tweet dopo aver effettuato la media degli score associati ai tweet emessi all'interno di ogni minuto.

Relativamente alla quantità di tweet emessi effettivamente all'interno di un minuto, di seguito vengono riportati i grafici delle criptovalute considerate nell'analisi nel periodo 27 settembre 2020 – 3 ottobre 2020.

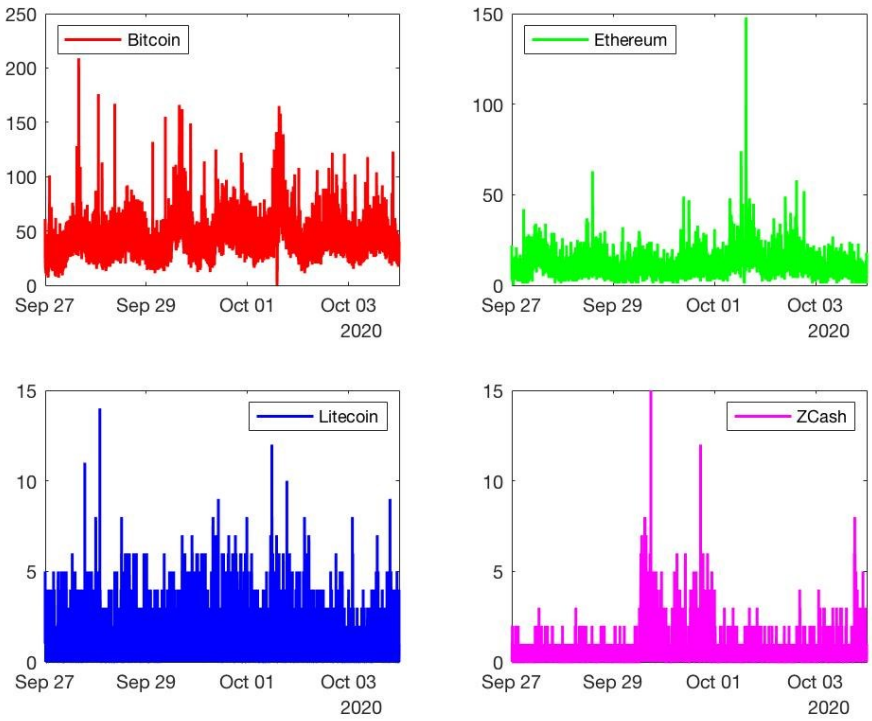

Figura 2.7 - Volumi dei tweet al minuto per le quattro criptovalute considerate.

Anche graficamente viene ribadita la proporzione del volume dei tweet emessi per le quattro criptovalute, la quale vede Bitcoin primeggiare sulle altre toccando il picco di circa 200 tweet in un minuto nel periodo d'analisi considerato.

Noti i dati finanziari sulle criptovalute relativi a prezzi di chiusura e volumi, gli score relativi ai tweet associati ad ogni minuto e il volume dei tweet per minuto è possibile procedere nell'analisi.

Di seguito viene mostrato il grafico rappresentante le fasi del processo di analisi seguito, dove alle fasi descritte, oltre a quella relativa alla raccolta dei dati si aggiunge quella relativa alla Sentiment Analysis.

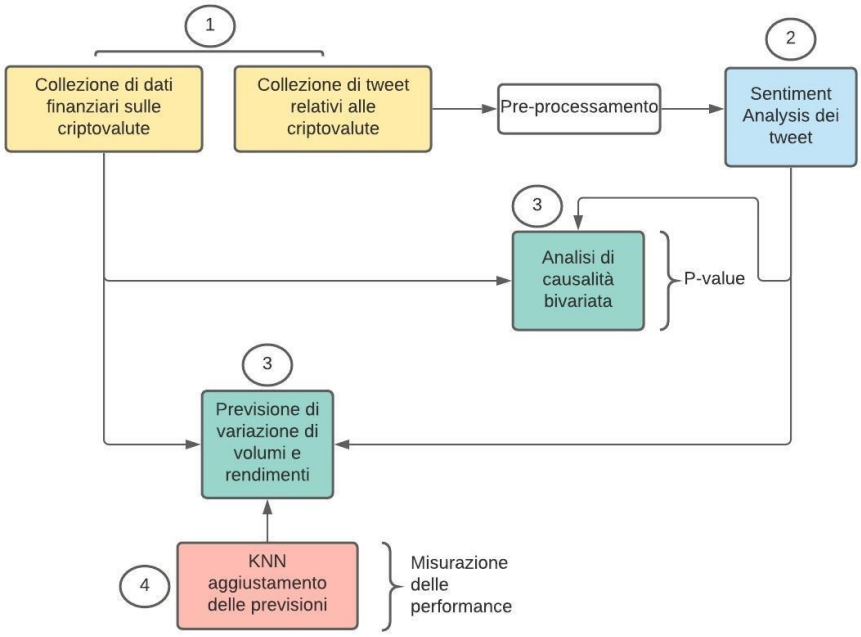

Figura 2.8 - Rappresentazione delle fasi del processo utilizzate per effettuare l'analisi.

Prima di passare a descrivere l'analisi di causalità e le previsioni delle variabili finanziarie effettuate tramite il VAR, è stato necessario applicare una trasformazione delle serie considerate. Ogni criptovaluta possiede quattro variabili: il prezzo di chiusura, i volumi, gli score dei tweet e i volumi dei tweet al minuto. Per ognuna di queste variabili è stata fatta una trasformazione per evitare problemi di stazionarietà all'interno dei modelli vettoriali autoregressivi VAR.

Inoltre, quello che viene considerato importante in questa sede non è il prezzo di chiusura o il volume in sé della singola criptovaluta, bensì la sua variazione da un periodo temporale all'altro, in questo caso il minuto.

Perciò, sono state applicate delle trasformazioni logaritmiche a tutte le serie storiche al fine di normalizzare e consentire il confronto tra le serie, allineandosi ai metodi utilizzati da Sprenger (2014) e Li (2017).

Di seguito vengono indicate le variabili rendimento R_t, variazione di volume V_t, score dei tweet S_t, e volumi dei tweet per minuto TV_t relative ad ognuna delle quattro criptovalute considerate, effettuando le seguenti trasformazioni:

$$R_t = \ln\left(\frac{p_t}{p_{t-1}}\right), \quad t \geq 1 \tag{2.1}$$

$$V_t = \ln\left(\frac{1+v_t}{1+v_{t-1}}\right), \quad t \geq 1 \tag{2.2}$$

$$S_t = \ln\left(\frac{1+s_t}{1+s_{t-1}}\right), \quad t \geq 1 \tag{2.3}$$

$$TV_t = \ln\left(\frac{1+tv_t}{1+tv_{t-1}}\right), \quad t \geq 1 \tag{2.4}$$

Il rendimento di una criptovaluta viene calcolato come logaritmo del rapporto tra il prezzo al periodo t e il prezzo al periodo $t-1$. Per le altre variabili considerate nelle analisi viene sommato 1 al numeratore e al denominatore per evitare dei

rapporti con denominatore nullo che avrebbero comportato errori durante le analisi svolte.

2.3.1 Analisi di causalità tra criptovalute e Twitter

Uno degli obiettivi di questa tesi è quello di analizzare circa il possibile nesso di causalità tra i dati finanziari delle criptovalute considerate e il sentimento ottenuto da Twitter riassumibile con il punteggio determinato attraverso VADER.

Per indagare rispetto alla causalità delle variabili sopra descritte è stato utilizzato un modello vettoriale autoregressivo VAR bivariato, in cui vengono analizzate le relazioni tra variabili di carattere finanziario e variabili relative al social media Twitter.

Attraverso l'analisi di causalità secondo Granger è stato possibile testare l'ipotesi nulla di non-granger causalità tra le due variabili considerate, ottenendo dei p-value che se tendenti allo zero ci permettono di rifiutare l'ipotesi nulla, affermando che tra le due esiste una relazione di causa-effetto unidirezionale.

Il test di causalità secondo Granger[23] è stato svolto considerando un modello vettoriale autoregressivo VAR di ordine *p* sottostante alle due variabili cui si vuole studiare la relazione, ed andando ad effettuare il test considerando ogni volta un VAR con ordine superiore fino all'ordine massimo $p = 15$.

[23] *Granger, "Investigating Causal Relations by Econometric Models and Cross-spectral Methods", Econometrica, Vol. 37, No. 3, pp. 424-438.*

Inoltre, la causalità è stata testata su tre differenti frequenze delle serie storiche di ogni criptovaluta, andando a considerare i valori ogni minuto, ogni dieci minuti, ogni venti minuti delle variabili rendimento R_t, variazione di volume V_t, score dei tweet S_t, e volumi dei tweet per minuto TV_t. In questa sede sono state studiate tutte le relazioni unidirezionali tra le variabili finanziarie (R_t, V_t) e le variabili dedotte dalla Sentiment Analysis sui tweet collezionati (S_t, TV_t).

2.3.2 Previsione dei dati finanziari mediante VAR

Un altro obiettivo della tesi è quello di effettuare delle previsioni ad alta frequenza delle variabili finanziarie, ovvero rendimenti e variazioni di volumi, delle criptovalute considerate.

Per effettuare previsioni è stato utilizzato un VAR bivariato di ordine p, con quest'ultimo selezionato attraverso il criterio di Akaike minore, inserendo una variabile finanziaria e una variabile maggiormente qualitativa derivante dal social media Twitter. In questo contesto sono stati preferibilmente utilizzati due accoppiamenti inseriti nel VAR bivariato, rendimenti e score (R_t, S_t), variazioni di volume e volumi di tweet (V_t, TV_t), riferiti alla singola criptovaluta per cui è stata effettuata la previsione della sola variabile finanziaria d'interesse.

Anche nell'effettuare previsioni delle variabili finanziarie riferite alle criptovalute sono state utilizzate tre diverse frequenze delle serie storiche di ciascuna, tenendo conto dei valori ogni minuto, ogni dieci minuti e ogni venti minuti.

Relativamente alle serie storiche con frequenza delle osservazioni di un minuto, sono state utilizzate le prime 300 osservazioni per effettuare previsioni con ampiezza di 1 periodo oppure di 6 periodi. Invece, per le serie storiche con frequenza delle osservazioni di 10 minuti oppure 20 minuti, sono state effettuate previsioni ad 1 periodo o a 6 periodi utilizzando le prime 100 osservazioni.

Quindi, il processo previsionale è stato ripetuto in entrambi i casi utilizzando una rolling window e riuscendo quindi ad ottenere una previsione per ogni osservazione disponibile nel set informativo.

Tuttavia, prima di procedere ad effettuare le previsioni è stato necessario individuare l'ordine ottimale relativo alla singola serie storica utilizzata per prevedere il valore futuro della variabile.

Ciò è stato possibile andando a determinare il criterio di Akaike per ogni ordine del VAR con $p = 1, \dots, 15$ ed andando ad individuare l'ordine p che rende il criterio d'informazione minimo, considerando l'ordine associato ottimale a rendere la previsione più accurata.

Attraverso le previsioni effettuate con il VAR bivariato, considerando ad esempio rendimenti e score relativi ad una criptovaluta, è possibile effettuare previsioni su entrambe le variabili, in cui ogni variabile viene utilizzata per prevedere l'altra oltre che sé stessa nel sistema considerato. Tuttavia, in questa sede si è concentrata l'attenzione solo alla previsione delle variabili finanziarie, migliorando quest'ultima con l'utilizzo delle variabili derivanti da Twitter.

Di seguito viene riportato il grafico delle previsioni relative a Bitcoin effettuate sui rendimenti R_t mediante VAR bivariato, utilizzando gli score S_t come altra variabile inserita nel modello, considerando la serie storica con frequenza di osservazioni di un minuto ed ampiezza previsionale pari a 1 periodo.

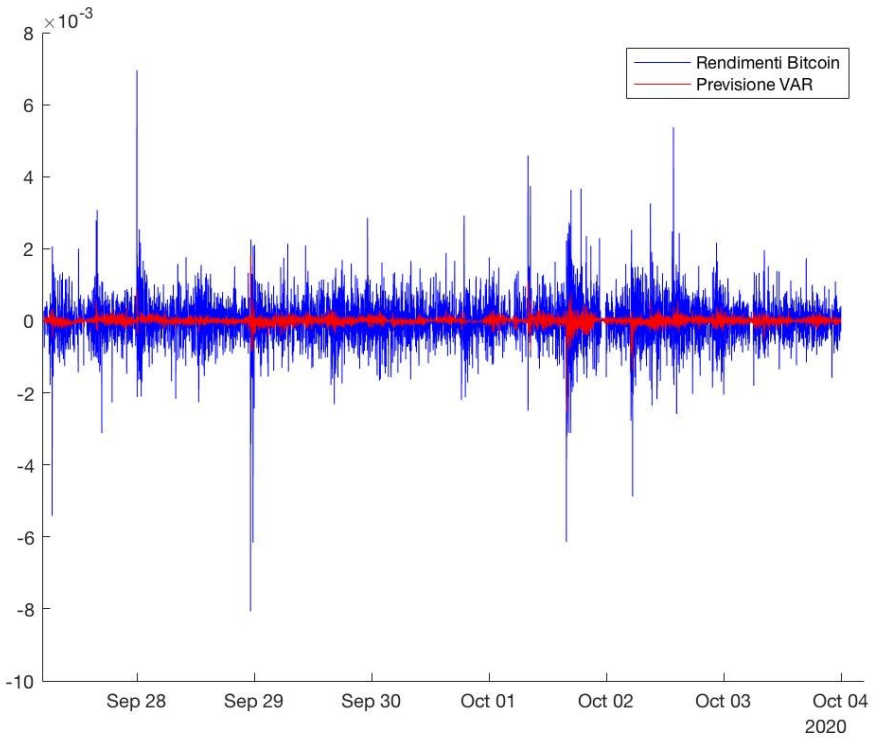

Figura 2.9 - Previsione del VAR e valori effettivi del rendimento di Bitcoin, frequenza osservazioni 1 minuto e ampiezza previsione 1 periodo..

Dalla Figura 2.9 si nota che i valori previsti dal VAR nel periodo 27 settembre 2020 – 3 ottobre 2020 non risultano essere relativamente accurati se confrontati con i valori reali del rendimento di Bitcoin relativi allo stesso periodo.

Tuttavia, il confronto tra i valori previsti dal VAR e i valori reali del rendimento di Bitcoin può essere visualizzato anche sotto forma di trend anziché attraverso la forma stazionaria delle due serie.

Infatti, assumendo $P_0 = P_{VAR,0} = 1$ e caricando i rendimenti attraverso la formula inversa, si possono ottenere i prezzi reali e i prezzi stimati dal VAR come mostrato in seguito:

$$P_{t+1} = P_t \cdot e^{R_t}, \qquad P_{VAR,t+1} = P_{VAR,t} \cdot e^{R_{t,VAR}} \tag{2.5}$$

dove, $R_{t,VAR}$ sono i rendimenti stimati dal VAR.

Di seguito vengono mostrate le previsioni effettuate dal VAR e i valori reali del rendimento di Bitcoin considerando la rappresentazione sotto forma di trend nel periodo 27 settembre 2020 – 3 ottobre 2020 per la serie storica con frequenza di osservazioni ad un minuto e ampiezza di previsione pari a 1 periodo.

Figura 2.10 - Previsione del VAR e valori effettivi del rendimento di Bitcoin frequenza osservazioni 1 minuto e ampiezza previsione 1 periodo (trend).

È possibile notare come il trend delle previsioni del VAR sembra seguire in maniera abbastanza accurata le variazioni di prezzo realmente subite da Bitcoin nel periodo analizzato. I rendimenti in questa analisi sono caratterizzati da oscillazioni di importo relativamente contenuto in quanto la previsione e i valori reali si riferiscono ai valori reali di Bitcoin in 1 minuto. Quindi, anche se, come ampiamente detto in precedenza, siamo di fronte a strumenti finanziari altamente volatili, le loro oscillazioni all'interno di un intervallo di un minuto risultano essere di dimensione relativamente ridotta.

2.3.3 Miglioramento delle previsioni attraverso KNN

A questo punto, dopo aver effettuato tutte le previsioni delle criptovalute analizzate per le tre frequenze delle serie storiche e per le due ampiezze di previsione considerate (1 periodo e 6 periodi), introduciamo il modello non parametrico K-Nearest-Neighbor (KNN) con l'obiettivo di migliorare le previsioni effettuate mediante VAR.

Infatti, in questo lavoro di tesi non verranno effettuate previsioni mediante KNN, bensì verrà sfruttato questo classificatore per migliorare le previsioni effettuate attraverso il modello vettoriale autoregressivo VAR. Quindi, una volta calcolate le previsioni effettuate dal VAR, si è proceduto a calcolare la differenza tra queste e i valori reali delle variabili finanziarie oggetto di analisi, come descritto dalle equazioni:

$$SC_t = R_{t,VAR} - R_t \quad o \quad SC_t = V_{t,VAR} - V_t \tag{2.6}$$

Perciò, la differenza tra queste due serie storiche (SC_t), è l'oggetto del classificatore K-Nearest-Neighbor e può essere denominato in questa sede scarto o residuo. Tuttavia, per evitare che valori anomali (outliers) potessero inficiare nella classificazione del vettore preso in esame, sono state applicate due tecniche di normalizzazione e pulizia dell'input da inserire nel classificatore. In particolare, è stato in precedenza eliminato il valore massimo e il valore minimo della serie degli

scarti, introducendo al loro posto la media tra il valore precedente e quello successivo. Poi, è stata effettuata la normalizzazione z-score citata nel primo capitolo, in cui al denominatore è stata utilizzata la deviazione media assoluta con lo scopo di attribuire ai valori anomali un minore impatto.

Gli aggiustamenti delle previsioni mediante KNN sono stati effettuati sulle ultime 50 previsioni ottenute con il VAR, utilizzando nel modello non parametrico un vettore di lunghezza *k,* con $k = 1, \dots, 50$.

Infatti, con il metodo KNN è stato valutato un vettore avente lunghezza *k* nella serie degli scarti descritta sopra, andando a determinare il vettore nel passato avente distanza euclidea minima da quello scelto in partenza e individuando il vettore di scarto successivo con lunghezza pari alle previsioni da aggiustare o correggere. Quindi, attraverso il calcolo della distanza euclidea per ciascun vettore avente lunghezza *k,* usando una rolling window e mantenendo *k* invariato, si è riuscito a determinare il vettore maggiormente simile a quello analizzato.

A questo punto, il vettore dei residui determinato è stato aggiunto alle stesse previsioni effettuate mediante il VAR ottenendo quindi delle previsioni aggiustate attraverso il KNN. Di seguito viene illustrato il grafico della serie degli scarti di Ethereum nel caso di previsioni mediante VAR a 6 periodi, analizzando la serie con frequenza di osservazioni a 10 minuti per le variazioni di volume V_t, scegliendo di correggere 20 previsioni con *k* pari 20.

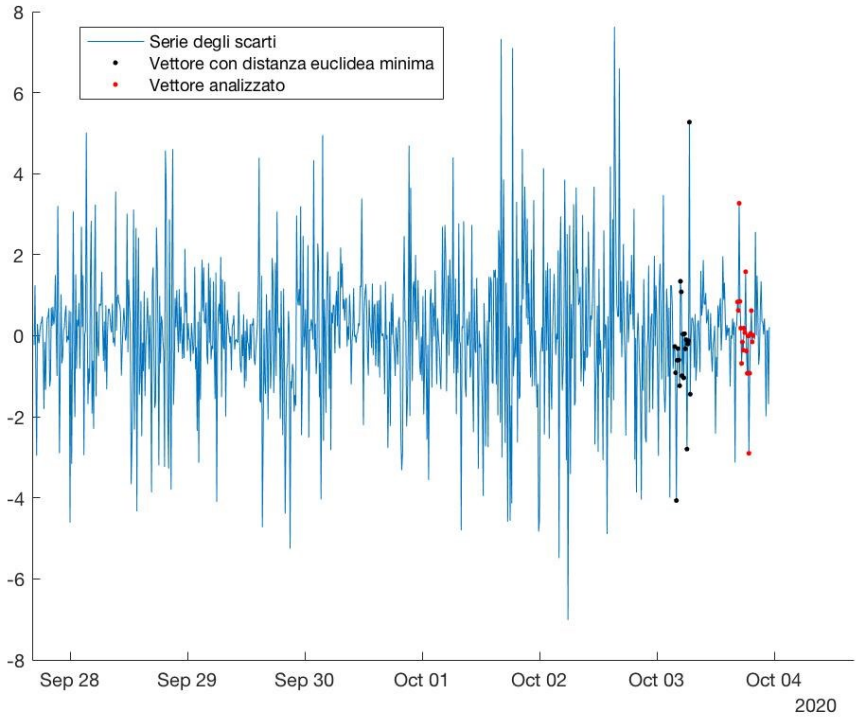

Figura 2.11 - Serie degli scarti SC_t delle previsioni delle variazioni di volume di Ethereum, frequenza osservazioni 10 minuti e ampiezza previsione 6 periodi.

I punti sottolineati in rosso sono relativi al vettore scelto con lunghezza $k = 20$, mentre i punti sottolineati in nero rappresentano il vettore selezionato lungo tutta la serie passata avente distanza euclidea minima dal vettore scelto.

Quindi, avendo scelto in questo caso di migliorare 20 previsioni effettuate con il VAR, a quest'ultime verranno aggiunti i 20 valori degli scarti successivi al vettore di punti sottolineati in nero rappresentato nella Figura 2.11.

La nuova previsione, data dalla somma della previsione effettuata con il VAR e il vettore degli scarti successivo a quello avente distanza euclidea minima, verrà denominata previsione del VAR aggiustata (R_{t,VAR_adj}).

2.3.4 Misurazione delle performance e scelta della previsione e *k* ottimali

In questo paragrafo confronteremo le previsioni ottenute con il VAR con quelle ottenute con il VAR aggiustate usando il KNN sui residui del VAR.

Per effettuare delle misurazioni dell'effettivo miglioramento delle previsioni in seguito all'adozione del classificatore K-Nearest-Neighbor sono stati considerati tre indici di performance che valutano in diverso modo quanto la previsione del VAR aggiustata risulti migliore della previsione del VAR originale. Inoltre, attraverso questi indici di performance è stato possibile individuare la previsione che ha ottenuto maggiori miglioramenti per qualsiasi valore di k utilizzato nel metodo KNN.

Se consideriamo che attraverso il KNN sono state corrette h^* previsioni, caricando attraverso la formula inversa le previsioni del VAR aggiustate e assumendo $P_{VAR\_adj,0} = 1$ si ottengono i prezzi aggiustati come riportato in seguito:

$$P_{VAR_{adj},t+1} = P_{VAR_{adj},t} \cdot e^{R_{VAR\_adj,t}} \tag{2.7}$$

Quindi, è possibile ottenere i residui dei prezzi del VAR aggiustato e i residui dei prezzi del VAR con le equazioni descritte in seguito:

$$SC_{VAR,t} = P_{VAR,t} - P_t, \qquad SC_{VAR\_adj,t} = P_{VAR\_adj,t} - P_t \tag{2.8}$$

dove P_t sono i prezzi realmente osservati delle criptovalute.

Perciò, in seguito alle considerazioni effettuate in precedenza è possibile descrivere l'indice di performance di distanza come segue:

$$Indice\ di\ performance\ di\ distanza = \frac{\left|SC_{VAR\_adj,t}\right| < \left|SC_{VAR,t}\right|}{h^*} \tag{2.9}$$

L'indice di performance descritto nell'Equazione 2.9 sfrutta il concetto di distanza tra vettori per misurare quanto la previsione del VAR aggiustata sia più vicina punto per punto ai valori reali della variabile, rispetto alla previsione del VAR. Infatti, per determinare l'indice sono state effettuate le differenze in valore assoluto tra la previsione del VAR aggiustata e i valori reali della variabile, e tra la previsione del VAR e i valori reali collezionati individuando in quanti punti sul totale delle previsioni corrette la prima differenza risulta minima all'altra. In questo caso, si considereranno come avvenuti dei miglioramenti nelle previsioni nel momento in cui l'indice di performance di distanza assume un valore maggiore o uguale al 50%.

Per quanto riguarda il secondo indice di performance determinato, questo tiene conto rispettivamente dei valori medi dei residui delle previsioni del VAR aggiustato e dei residui delle previsioni del VAR. Di seguito viene descritta l'equazione utilizzata per determinare l'indice di performance in media:

$$Indice\ di\ performance\ in\ media = \frac{\mu_{SC_{VAR\_adj}} - \mu_{SC_{VAR}}}{\mu_{SC_{VAR}}} \qquad (2.10)$$

dove $\mu_{SC_{VAR\_adj}}$ e $\mu_{SC_{VAR}}$ rappresentano rispettivamente i valori medi dei residui delle previsioni del VAR aggiustato e dei residui delle previsioni del VAR. Per come è stato costruito l'indice di performance in media, si può segnalare un miglioramento nel momento in cui l'indice assume segno negativo.

Infine, il terzo indice di performance considera la somma dei quadrati dei residui rispettivamente della previsione aggiustata del VAR e della previsione ottenuta mediante il VAR. Di seguito viene descritta l'equazione utilizzata nel calcolo dell'indice di performance dei residui:

$$Indice\ di\ performance\ dei\ residui = \frac{\sum {SC_{VAR\_adj,t}}^2 - \sum {SC_{VAR,t}}^2}{\sum {SC_{VAR,t}}^2} \qquad (2.11)$$

Anche nel caso dell'indice di performance dei residui, nella discussione dei risultati ottenuti si farà menzione di miglioramenti qualora l'indice in questione assuma valori negativi.

Tuttavia, per poter determinare la lunghezza *k* ottimale da utilizzare per migliorare le previsioni di una determinata lunghezza, è stata effettuata un'analisi di robustezza. Infatti, come descritto nel capitolo precedente, non esiste una regola precisa per la selezione del valore di *k* maggiormente appropriato, pertanto questo deve essere individuato empiricamente.

Quindi, per misurare la robustezza dei risultati ottenuti e determinare il *k* migliore con cui aggiustare la previsione, è stata scelta la lunghezza previsionale per la quale la frequenza dei miglioramenti raggiunge il valore massimo.

A questo punto, scelta la lunghezza della previsione da migliorare, sono stati calcolati, con i metodi descritti in precedenza, gli indici di performance per ogni *k* per le 50 serie passate, le quali vengono ottenute eliminando dalla serie di partenza ogni volta l'ultima osservazione. Infine, per ogni *k* valutato sulla previsione ottimale scelta in precedenza, è stata determinata la frequenza dei miglioramenti ottenuti nelle 50 serie passate, individuando il *k* avente frequenza di miglioramenti massima. Perciò, la frequenza dei miglioramenti ottenuti nelle 50 serie passate è attribuibile alla robustezza del risultato ottenuto nell'indice di performance migliorando le previsioni con la combinazione di previsione e *k* ottimi.

Capitolo 3

RISULTATI EMPIRICI DELLE ANALISI EFFETTUATE

3.1 RISULTATI DERIVANTI DALL'ANALISI DI CAUSALITÀ

Uno dei principali obiettivi di questa tesi è quello di individuare una possibile relazione tra il mercato delle criptovalute e il social media Twitter.

A tal proposito è stato utilizzato un modello VAR bivariato in cui sono state inserite le variabili rendimenti R_t, variazioni di volume V_t relative alle criptovalute considerate e le variabili score dei tweet S_t, volume dei tweet TV_t relative a Twitter. Perciò, attraverso il test di Granger è stato possibile ottenere dei risultati circa tutte le relazioni possibili tra le variabili finanziarie e le variabili derivanti da Twitter.

Quindi, è stato effettuato un test di Granger unidirezionale per le quattro variabili sopra citate in modo da andare ad analizzare se una variabile "granger-causa" o meno l'altra variabile e viceversa. Il test di Granger relativo a ciascuna relazione è stato ripetuto utilizzando ogni volta un modello autoregressivo vettoriale VAR con ordine crescente $p = 1, \dots, 15$ e andando a considerare per ogni criptovaluta le serie storiche aventi tre diverse frequenze: 1 minuto, 10 minuti, 20 minuti.

Infine, sono stati ottenuti i p-value relativi a ciascuna relazione delle variabili considerate in cui è stato utilizzato un modello VAR con differenti ritardi. In

Appendice sono riportate le Tabelle 1-12 relative ai risultati ottenuti per le quattro criptovalute considerate, nelle diverse frequenze di dati analizzate.

In questo contesto è importante soffermarsi ed individuare i valori dei p-value prossimi allo zero poiché questi comportano una maggiore probabilità di rifiutare l'ipotesi nulla di non Granger-causalità. Quindi nel prosieguo della discussione dei risultati ottenuti si farà particolare attenzione a quelle relazioni il cui p-value è inferiore a 0,10, 0,05, oppure inferiore a 0,01 come sottolineato nelle tabelle riportate nell'Appendice.

Infatti, per quanto riguarda Bitcoin, considerando la frequenza di osservazione ad 1 minuto (Tabella A.1), è stata individuata una relazione di Granger-causalità tra il volume dei tweet TV_t e i rendimenti R_t. Infatti, per i ritardi 8,9,10,11,14 il p-value del test di Granger è inferiore a 0,10 mentre per i ritardi 12,13 e 15 il p-value risulta essere inferiore a 0,05 indicando che il volume dei tweet "causa" i rendimenti riferiti a Bitcoin per la frequenza ad 1 minuto.

Relativamente ai dati di Bitcoin con frequenza 10 minuti (Tabella A.2), si sono individuate relazioni di Granger causalità relativamente alle variabili TV_t-V_t, S_t-V_t, TV_t-R_t in cui l'ordine implica che la prima variabile causa la seconda. Infatti, relativamente ai ritardi 4,5,6 è stato individuato che il volume dei tweet TV_t causi la variazione dei volumi V_t, poiché il p-value risulta essere inferiore a 0,10. Quest'ultima inoltre è stata individuata essere causata dallo score relativo ai tweet S_t per i ritardi 7,8,10,11 (p-value<0,10) e per i ritardi 2,4,5,9 (p-value<0,05).

Invece, riguardo alla relazione tra il volume dei tweet TV_t e i rendimenti R_t è stata individuata una causalità con p-value inferiore a 0,10 per l'ordine del VAR pari a 1. Infine, relativamente alla frequenza di osservazioni di 20 minuti per Bitcoin (Tabella A.3), sono state individuate relazioni di causalità con p-value inferiore a 0,10 al ritardo 1 per le coppie di variabili V_t-TV_t e V_t-S_t, in cui la variazione di volume causa lo score e il volume dei tweet associati a Bitcoin.

Per quanto riguarda la seconda criptovaluta più capitalizzata, Ethereum, andando a valutare i risultati per le osservazioni con frequenza pari a 1 minuto (Tabella A.4), si può osservare come la Granger causalità sia individuabile al ritardo 1 (p-value<0,10) e al ritardo 2 (p-value<0,05) nella relazione rendimenti R_t e score dei tweet S_t. Invece, nelle osservazioni di Ethereum con frequenza ogni 10 minuti (Tabella A.5), è osservabile come il volume dei tweet TV_t "granger causi" la variazione di volume V_t ai ritardi 3,4,5,6,7,8,9, con p-value inferiore a 0,05 e ai ritardi 10,11,13,15 con p-value inferiore a 0,10. Infine, nei dati di Ethereum osservati ogni 20 minuti (tabella A.6), si può sottolineare come la relazione di causa effetto tra TV_t-V_t e V_t-TV_t sia bidirezionale e con p-value inferiore a 0,05.

In relazione ai dati di Litecoin con frequenza di osservazione pari a 1 minuto (Tabella A.7), in particolare, si possono notare delle relazioni di causalità tra volumi dei tweet TV_t e variazioni di volume V_t, in cui risultano esserci p-value inferiori a 0,10 per i ritardi 11,15 e p-value inferiori a 0,05 per i ritardi 13,14. Inoltre, si è notata una relazione di causa-effetto sempre per la variazione di volume V_t, ma in

questo caso la variabile che "granger causa" quest'ultima risulta essere lo score dei tweet S_t. In quest'ultimo caso, sono stati riscontrati p-value inferiori a 0,10 per i ritardi 7,13,14,15 e p-value inferiori a 0,05 per i ritardi 10,11,12.

Per le osservazioni con frequenza di 10 minuti della criptovaluta Litecoin (Tabella A.8), è stato riscontrato un'ottimo risultato, unico in tutti i test di Granger effettuati. Infatti, al ritardo 1 è stato riscontrato che lo score dei tweet S_t risulta "granger-causare" i rendimenti R_t (p-value<0,05), permettendo al risultato di allinearsi con quelli ottenuti da Kraaijeveld, De Smedt (2020) sempre per Litecoin ma con frequenza delle osservazioni giornaliera.

Invece, per le ossservazioni di Litecoin con frequenza pari a 20 minuti (Tabella A.9), è sottolineabile come vi sia una relazione di causalità tra le variazioni di volumi V_t e gli score dei tweet S_t con p-value inferiore a 0,05 nei ritardi 2,3,4 e p-value inferiore a 0,10 al ritardo 5. Inoltre, è stato anche riscontrato che i rendimenti di Litecoin causino il volume dei tweet con p-value inferiore a 0,05 al ritardo 7 e con p-value inferiore a 0,10 al ritardo 8.

Relativamente ai risultati ottenuti per ZCash, si può notare che, nella frequenza di osservazioni ad 1 minuto (Tabella A.10), sono le variabili finanziarie R_t e V_t a causare rispettivamente il volume dei tweet TV_t (p-value<0,10) ai ritardi 1,2 e lo score dei tweet S_t (p-value<0,05) al ritardo 1.

Per quanto riguarda le osservazioni di ZCash con frequenza 10 minuti (Tabella A.11), è da segnalare come lo score dei tweet S_t "granger-causi" la variazione dei

volumi V_t ai ritardi 3,10,11 (p-value<0,10), ai ritardi 5,7,8,9 (p-value<0,05) e ai ritardi 6,12,13,14,15 (p-value<0,01). Inoltre, in questa frequenza di osservazione è da segnalare un'altra relazione causa-effetto significativa, in cui la variazione di volume V_t causa il volume dei tweet TV_t ai ritardi 4,9,10,11 (p-value<0,10). Nella stessa frequenza di osservazione si nota inoltre una relazione di causa effetto tra TV_t e V_t con p-value inferiore a 0,10 al ritardo 3.

Infine, relativamente alla frequenza di osservazione a 20 minuti per ZCash (Tabella A.12), c'è da segnalare come la variazione di volumi V_t "granger causi" il volume dei tweet TV_t ai ritardi 1,7,13,15 (p-value<0,10) e al ritardo 14 (p-value<0,05).

Quindi, in conclusione si può affermare che, anche se per un periodo dell'analisi non apparentemente ampio, esistono relazioni di causa-effetto secondo il test di Granger tra le variabili finanziarie delle criptovalute considerate e il social media Twitter. Infatti, come riscontrato nelle analisi effettuate, in alcuni casi sono le variabili finanziarie a causare gli score dei tweet e il volume dei tweet emessi, in altri, invece, di maggiore interesse, sono le variabili relative a Twitter ad influenzare il mercato delle criptovalute considerate. In particolare, l'obiettivo principale era quello di ricercare una relazione causale tra gli score dei tweet, che rappresentano le opinioni e il sentimento degli operatori, e i rendimenti e le variazioni di volume delle valute digitali. Questo è stato individuato nella criptovaluta Litecoin, in cui nelle osservazioni con frequenza a 10 minuti, si è riscontrato che gli score dei tweet causino i rendimenti di Litecoin con un'ottima significatività.

3.2 RISULTATI DEI MIGLIORAMENTI DELLE PREVISIONI DEL VAR

L'altro obiettivo prefissato in questa tesi è quello di effettuare attraverso il VAR delle previsioni circa le variabili finanziarie relative alle criptovalute considerate e, migliorarle mediante il classificatore KNN.

Come già detto nel precedente capitolo, i risultati sono stati valutati attraverso tre indici di performance, i quali valutano secondo diversi aspetti se è stata migliorata la previsione dei rendimenti o delle variazioni di volume delle valute digitali. Quindi, gli indici di performance di distanza, media e dei residui sono stati calcolati per tutte le combinazioni di previsione e k ottenendo delle valutazioni per ogni frequenza di osservazione.

Il primo passo è stato quello di individuare la previsione $(1, \dots, 50)$ con totale miglioramenti massimo, cercando di selezionare una previsione con periodo previsionale più ampio. In seconda battuta, è stata effettuata un'analisi di robustezza sulle 50 serie passate per andare a scegliere il valore di k $(1, \dots, 50)$, in cui sono avvenuti più miglioramenti, per effettuare la correzione della previsione ottenuta con il VAR. Quindi così facendo è stata individuata una combinazione previsione-k che si riferisce ad un risultato ottenuto per i tre indici di performance con una relativa robustezza. Di seguito vengono riportati i grafici ad istogramma delle criptovalute relativamente agli indici di performance in cui sono avvenuti maggiori miglioramenti, andando a segnalare la previsione scelta avente totale miglioramenti massimo per ogni k. Il primo indice di perfomance, che chiameremo indice di

distanza, analizzato misura la percentuale di volte in cui il VAR "aggiustato" con l'uso del machine learning fornisce una previsione più vicina a quella reale della previsione ottenuta con il VAR.

La prima figura fa riferimento alle variazioni di volume di Ethereum con frequenza di osservazione a 10 minuti, in cui sono state corrette le previsioni effettuate con il VAR aventi ampiezza pari a 1 periodo, misurate con l'indice di performance di distanza.

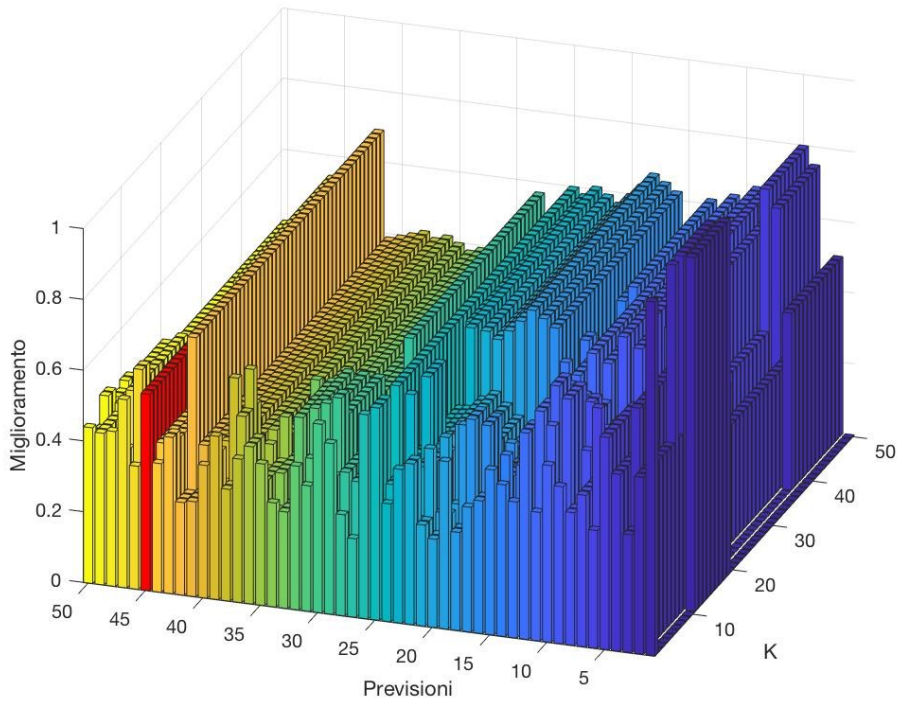

Figura 3.1 - Indice di performance di distanza, V_t Ethereum, frequenza osservazioni 10 minuti, ampiezza previsione 1 periodo.

Come si può notare dalla Figura 3.1, sono numerose le combinazioni di previsione e *k* in cui l'indice di distanza risulta essere maggiore o uguale al 50%, stando a significare che il 50% o più delle previsioni corrette con il KNN risulta essere meno distante dai valori reali rispetto alle previsioni del VAR. In questo specifico caso è stata scelta la previsione a 45 periodi, evidenziata in rosso, poiché possiede dei miglioramenti del 56% per tutti i valori di *k.*

Il secondo indice di performance considerato, che chiameremo indice in media, misura la differenza relativa tra la media dei residui associati alle previsioni aggiustate con il KNN e quella dei residui associati alle previsioni del VAR. Relativamente all'indice di performance in media (Equazione 2.10) si fa riferimento a Ethereum con frequenza di osservazioni a 20 minuti, in cui sono state corrette le previsioni del VAR aventi ampiezza previsionale di 6 periodi.

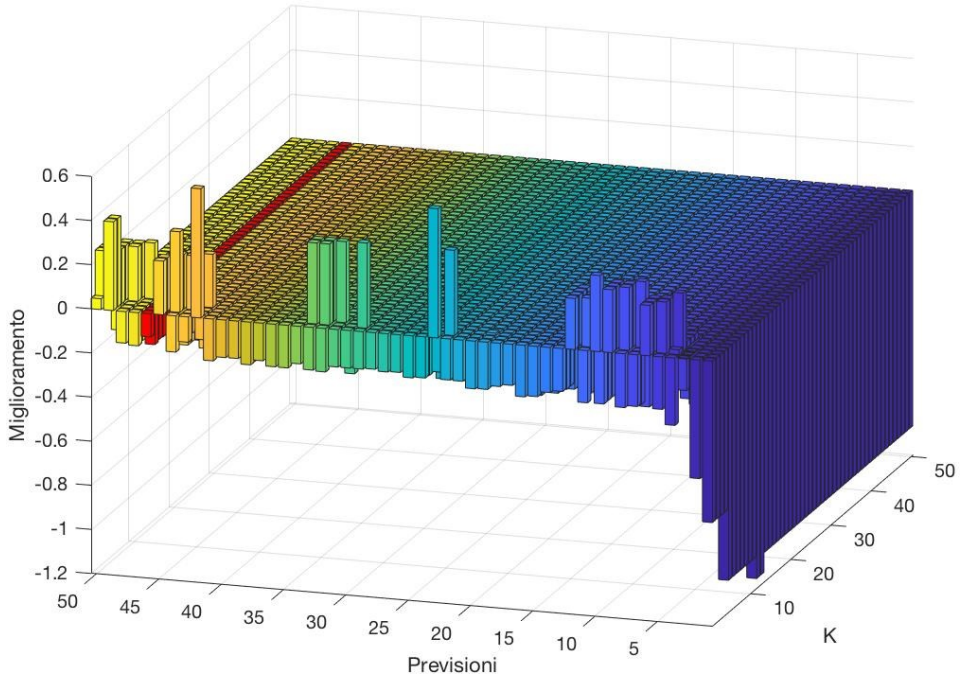

Figura 3.2 - Indice di performance in media, V_t Ethereum, frequenza osservazioni 20 minuti, ampiezza previsione 6 periodi.

In questo caso, sono stati ottenuti valori negativi dell'indice di performance ma questo, per come è stato costruito l'indice di performance in media, dimostra che le previsioni aggiustate con il KNN hanno in media dei residui minori rispetto alle previsioni del VAR. Anche per l'indice di performance in media è stata individuata la previsione avente totale miglioramenti massimo, nel caso del grafico precedente pari a 46 periodi (istogrammi evidenziati in rosso).

Il terzo e ultimo indice considerato, che chiameremo indice dei residui, misura la differenza relativa tra la somma dei quadrati dei residui delle previsioni aggiustate con il KNN e la somma dei quadrati dei residui delle previsioni ottenute col VAR. Per quanto riguarda l'indice di performance dei residui il totale dei miglioramenti massimo è stato individuato anche questa volta nelle variazioni di volumi di Ethereum con frequenza 20 minuti e ampiezza previsionale a 6 periodi.

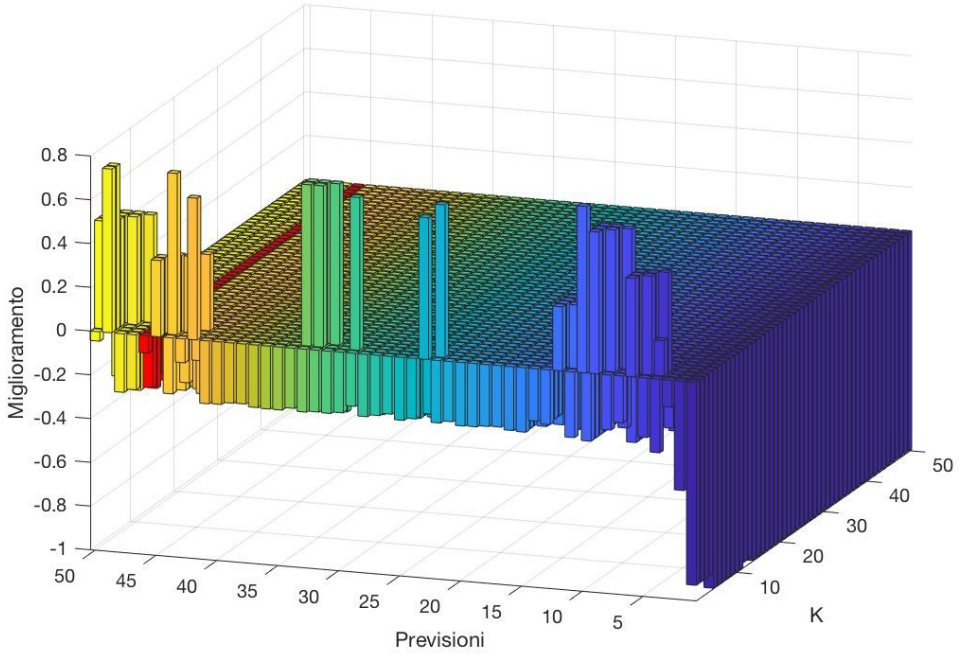

Figura 3.3 - Indice di performance dei residui, V_t Ethereum, frequenza osservazioni 20 minuti, ampiezza previsione 6 periodi.

Anche in questo caso i risultati ottenuti possiedono segno negativo, tuttavia, sono riconducibili a dei miglioramenti, in quanto, dovutamente alla costruzione

dell'indice di performance dei residui, dei valori negativi stanno ad indicare che la somma dei quadrati dei residui delle previsioni aggiustate con il KNN è inferiore rispetto a quella delle previsioni del VAR.

Persino per gli indici di performance dei residui è stata scelta la previsione avente totale miglioramenti massimo, che nella Figura 3.3 risulta essere quella riferita al periodo 46 (istogrammi evidenziati in rosso).

Nel prosieguo della trattazione verranno mostrate le migliori combinazioni di previsione-k per i singoli indici di performance mettendoli a confronto con il risultato ottenuto utilizzando la stessa combinazione ma con un altro indice di performance.

3.2.1 Miglioramenti nei rendimenti: ampiezza previsionale 1 periodo

Come descritto in precedenza le previsioni sulle variabili finanziarie (rendimenti, variazioni di volumi) sono state effettuate mediante il VAR su due diversi periodi previsionali: 1 periodo e 6 periodi, e poi corrette attraverso il KNN.

In questa sezione verranno trattati i risultati ottenuti dei tre indici di performance per le frequenze di osservazione dei dati delle criptovalute considerate relativamente ai rendimenti R_t per l'ampiezza previsionale ad 1 periodo.

Iniziando a discutere i risultati delle migliori combinazioni di previsione-k individuate mediante indice di performance di distanza, di seguito viene riportata

la tabella associata ai miglioramenti ottenuti per le diverse frequenze di osservazione relativamente ai rendimenti R_t delle quattro criptovalute.

Tabella 3.1 - Indice performance di distanza, rendimenti R_t, ampiezza previsionale 1 periodo.

		INDICE PERFORMANCE DISTANZA				INDICE PERFORMANCE RESIDUI	INDICE PERFORMANCE MEDIA
	FREQUENZA OSSERVAZIONI	PREVISIONI	K	% MIGLIORAMENTO	ROBUSTEZZA	% MIGLIORAMENTO	% MIGLIORAMENTO
BTC	1 minuto	2	25	50%	70%	94,47%	1035,75%
	10 minuti	16	6	50%	8%	48,68%	-132,17%
	20 minuti	**2**	**30**	**50%**	**66%**	**-61,11%**	**-34,77%**
ETH	1 minuto	2	34	50%	66%	295,93%	157,06%
	10 minuti	4	5	50%	34%	115,25%	-29,68%
	20 minuti	2	20	50%	46%	734,11%	118,72%
LTC	1 minuto	38	45	53%	38%	42,94%	-46,69%
	10 minuti	17	39	71%	16%	1,91%	-110,97%
	20 minuti	2	37	50%	24%	118,90%	30,59%
ZC	1 minuto	44	29	55%	38%	22,06%	-6,29%
	10 minuti	2	15	50%	28%	41,40%	-14,34%
	20 minuti	46	27	72%	18%	20,85%	-79,65%

Dalla Tabella 3.1 si osserva che per Bitcoin con frequenza di osservazioni a 20 minuti, se viene fatta una previsione a 2 periodi utilizzando nel metodo KNN un k pari a 30, l'indice di performance di distanza ci indica che vengono migliorate il 50% delle previsioni con una robustezza nelle 50 serie passate del 66%. Se presa la stessa combinazione di previsione-k (2-30) l'indice di performance dei residui ci indica che la previsione del VAR risulta essere peggiore della previsione aggiustata

del VAR del 61,11% e per l'indice di performance in media, le previsioni del VAR sono peggiori delle previsioni aggiustate del VAR del 34,77%.

Di seguito viene mostrato il grafico relativo al caso sottolineato nella Tabella 3.1, in cui sono state migliorate 2 previsioni per Bitcoin nelle osservazioni con frequenza 20 minuti.

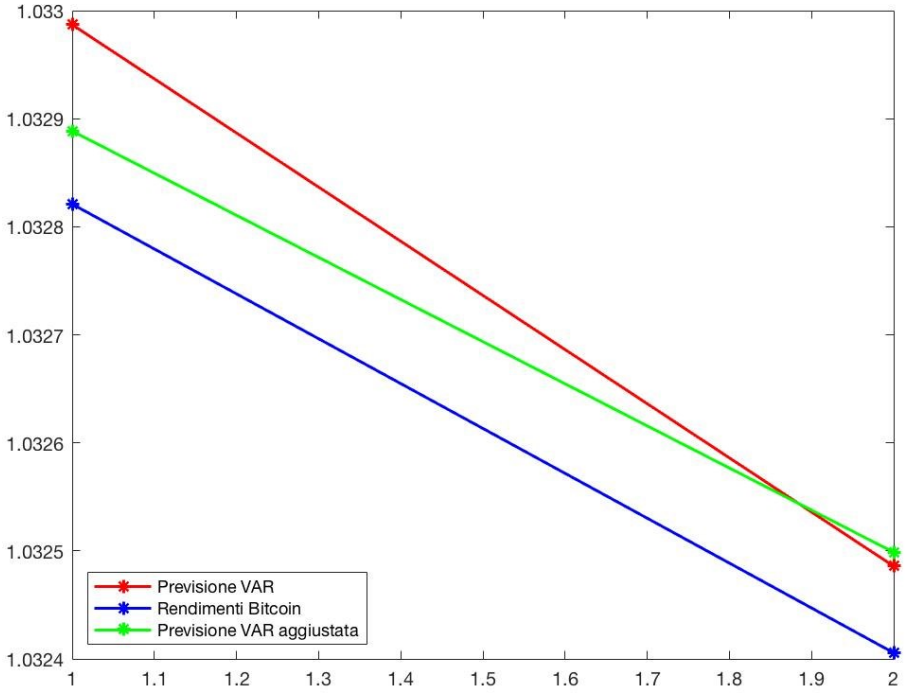

Figura 3.4 - Indice di performance di distanza, rendimenti R_t Bitcoin, frequenza osservazioni 20 minuti.

Come si può notare dalla Figura 3.4 il miglioramento valutato tramite la distanza delle previsioni dai valori reali è del 50%, sottolineando come il miglioramento abbia mantenuto il trend delle previsioni effettuate con il VAR ma avvicinato la prima previsione al valore realmente osservato. Scegliendo le combinazioni di

previsione-k per i rendimenti R_t delle quattro criptovalute considerate valutando in partenza l'indice di performance in media sono stati ottenuti i risultati riassunti nella seguente tabella.

Tabella 3.2 - Indice di performance in media, rendimenti R_t, ampiezza previsionale 1 periodo.

		INDICE PERFORMANCE MEDIA				INDICE PERFORMANCE RESIDUI	INDICE PERFORMANCE DISTANZA
	FREQUENZA OSSERVAZIONI	PREVISIONE	K	% MIGLIORAMENTO	ROBUSTEZZA	% MIGLIORAMENTO	% MIGLIORAMENTO
BTC	1 minuto	25	3	-1386,00%	62%	879,71%	16%
	10 minuti	10	48	-123,90%	56%	112,23%	40%
	20 minuti	5	34	-0,43%	56%	16,16%	40%
ETH	1 minuto	49	3	-11,19%	62%	202,27%	20%
	10 minuti	**49**	**50**	**-1595,34%**	**94%**	**231,61%**	**27%**
	20 minuti	48	35	-50,44%	62%	246,09%	21%
LTC	1 minuto	32	13	-71,12%	68%	71,13%	41%
	10 minuti	32	50	-112,99%	54%	189,55%	44%
	20 minuti	50	33	-18,82%	62%	243,57%	20%
ZC	1 minuto	44	27	-6,29%	74%	22,06%	55%
	10 minuti	30	6	-133,03%	64%	104,63%	53%
	20 minuti	50	50	-266,30%	66%	62,91%	34%

In questo caso, scegliendo le combinazioni di previsione-k partendo dall'indice di performance in media con il processo illustrato in precedenza, non si ottengono miglioramenti con risultati ottimi in ciascuno degli indici valutati. Tuttavia, in Ethereum con frequenza di osservazione a 10 minuti per la combinazione di previsione-k, 49-50, le previsioni del VAR aggiustato sono in media migliori del

1595,34% rispetto alle previsioni effettuate con il VAR. Di seguito viene riportato il grafico dei miglioramenti con previsione pari a 49 e *k* uguale a 50 per i rendimenti di Ethereum con frequenza di osservazioni a 10 minuti.

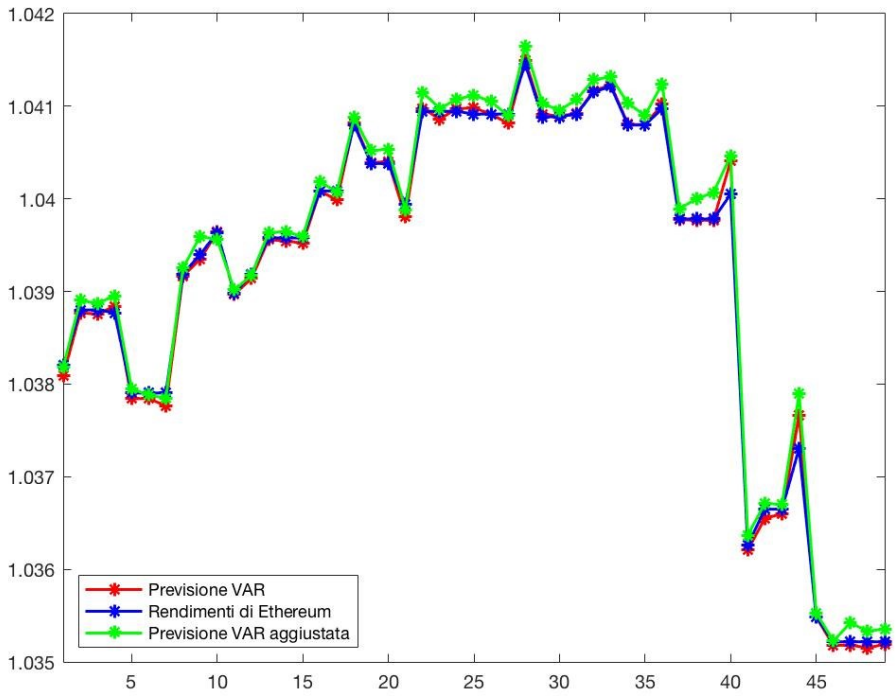

Figura 3.5 - Indice di performance in media, rendimenti R_t Ethereum, frequenza osservazioni 10 minuti.

A primo impatto non si può dire che le previsioni del VAR siano state migliorate poiché, come riportato dall'indice di performance di distanza solo 14 previsioni aggiustate hanno distanza minore rispetto alle previsioni effettuate dal VAR, tuttavia, in media le previsioni aggiustate del VAR risultano migliori del 1595,34 %.

Per quanto riguarda l'indice di performance dei residui, di seguito viene riportata la tabella che illustra i risultati ottenuti valutando inizialmente quest'ultimo indice per scegliere le combinazioni di previsione-*k* ottime delle quattro criptovalute analizzate.

Tabella 3.3 - Indice di performance dei residui, rendimenti R_t, ampiezza previsionale 1 periodo.

		INDICE PERFORMANCE RESIDUI				INDICE PERFORMANCE DISTANZA	INDICE PERFORMANCE MEDIA
	FREQUENZA OSSERVAZIONI	PREVISIONI	K	% MIGLIORAMENTO	ROBUSTEZZA	% MIGLIORAMENTO	% MIGLIORAMENTO
BTC	1 minuto	1	10	-5,90%	46%	100%	-2,99%
	10 minuti	12	6	-10,11%	4%	58%	-33,25%
	20 minuti	**2**	**30**	**-61,11%**	**66%**	**50%**	**-34,77%**
ETH	1 minuto	3	49	-8,99%	20%	33%	-23,06%
	10 minuti	16	16	-9,94%	8%	44%	6,60%
	20 minuti	1	12	-44,47%	32%	100%	-25,48%
LTC	1 minuto	5	17	-33,74%	34%	80%	-7,64%
	10 minuti	10	33	-48,97%	12%	80%	-96,32%
	20 minuti	20	-	-	-	-	-
ZC	1 minuto	29	47	-4,47%	4%	86%	-0,47%
	10 minuti	1	15	-99,90%	20%	100%	-96,79%
	20 minuti	9	3	-16,81%	8%	78%	-38,00%

Anche in questo caso, come nei risultati dell'indice di performance di distanza, è sottolineabile il miglioramento avvenuto in Bitcoin con frequenza di osservazione a 20 minuti nella combinazione di previsione-*k* 2-30 avendo una robustezza dei risultati del 66%. Tuttavia, è da segnalare il caso di Litecoin con frequenza di

osservazione a 20 minuti, in cui era stata individuata la previsione a 20 periodi futuri ma effettuando l'analisi di robustezza non sono stati trovati miglioramenti nelle 50 serie precedenti per qualsiasi *k*.

3.2.2 Miglioramenti nei volumi: ampiezza previsionale 1 periodo

Per quanto riguarda le variazioni di volumi anche in questo caso verranno illustrati i risultati ottenuti, individuando in ogni singolo indice di performance la combinazione ottima di previsione e *k* e verrà confrontata con i risultati degli altri indici.

Scegliendo la previsione ottima e il relativo *k* dai risultati ottenuti per l'indice di performance di distanza sono stati conseguiti i miglioramenti descritti nella tabella riportata di seguito.

Tabella 3.4 - Indice di performance di distanza, volumi V_t, ampiezza previsionale 1 periodo.

		INDICE PERFORMANCE DISTANZA				INDICE PERFORMANCE RESIDUI	INDICE PERFORMANCE MEDIA
	FREQUENZA OSSERVAZIONI	PREVISIONI	K	% MIGLIORAMENTO	ROBUSTEZZA	% MIGLIORAMENTO	% MIGLIORAMENTO
BTC	1 minuto	2	11	50%	82%	-7,40%	-24,77%
	10 minuti	33	30	55%	8%	15,06%	62,95%
	20 minuti	2	44	50%	72%	-13,10%	161,56%
ETH	1 minuto	2	50	50%	54%	-21,74%	-57,97%
	10 minuti	45	23	56%	22%	368,87%	25,56%
	20 minuti	2	26	50%	76%	15388,68%	-886,48%
LTC	1 minuto	2	9	50%	56%	16,38%	-86,12%
	10 minuti	4	26	50%	48%	41,74%	158,46%
	20 minuti	3	28	100%	24%	-37,82%	-20,06%
ZC	**1 minuto**	**2**	**8**	**100%**	**54%**	**-31,23%**	**-56,31%**
	10 minuti	2	49	50%	70%	-23,69%	-12,07%
	20 minuti	2	9	50%	72%	0,80%	-8,90%

Nei risultati mostrati nella Tabella 3.4 è da sottolineare il miglioramento riscontrato in ZCash con frequenza di osservazione dei dati ad 1 minuto. Di seguito viene mostrato il grafico delle previsioni di ZCash migliorate al 100% attraverso la combinazione di previsione pari a 2 e *k* pari a 8 secondo l'indice di performance di distanza.

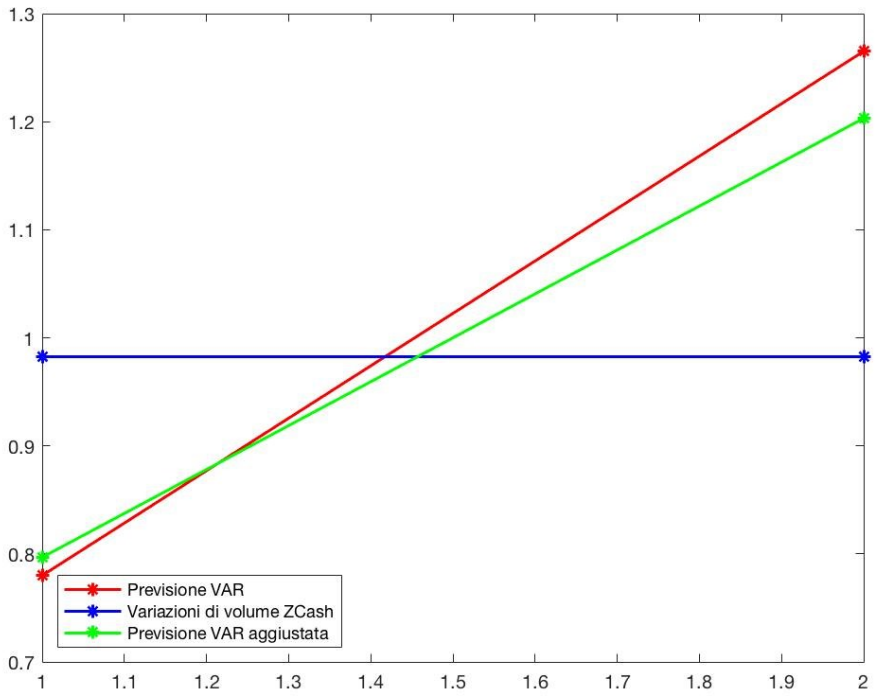

Figura 3.6 - Indice di performance di distanza, volumi V_t ZCash, frequenza osservazioni 1 minuto.

Come si nota dalla Figura 3.6 le previsioni del VAR aggiustate possiedono una distanza minore dai valori realmente osservati rispetto alle previsioni del VAR, riscontrando in questo caso anche un miglioramento nell'indice di performance in media e dei residui secondo cui il VAR performa peggio del VAR aggiustato mediante KNN. Relativamente alla scelta delle previsioni da migliorare e del k ottimo associato individuati mediante l'indice di performance in media, sono stati ottenuti i seguenti risultati per le variazioni di volume delle criptovalute analizzate.

Tabella 3.5 - Indice di performance in media, volumi V_t, ampiezza previsionale 1 periodo.

		INDICE PERFORMANCE MEDIA				INDICE PERFORMANCE RESIDUI	INDICE PERFORMANCE DISTANZA
	FREQUENZA OSSERVAZIONI	PREVISIONI	K	% MIGLIORAMENTO	ROBUSTEZZA	% MIGLIORAMENTO	% MIGLIORAMENTO
BTC	1 minuto	5	45	-48,57%	52%	162,73%	20%
	10 minuti	2	10	-12,11%	34%	-56,85%	50%
	20 minuti	7	49	-64,25%	18%	-34,54%	71%
ETH	1 minuto	14	13	-470,73%	58%	14,45%	36%
	10 minuti	6	2	-10576,92%	36%	708,28%	67%
	20 minuti	5	50	-2444,64%	44%	10323,69%	40%
LTC	1 minuto	2	5	-2210,09%	44%	25385,83%	0%
	10 minuti	8	38	-950101,81%	60%	42449026,82%	0%
	20 minuti	42	50	-763714,29%	38%	500690814,7%	0%
ZC	1 minuto	3	16	-6,19%	44%	-1,88%	100%
	10 minuti	**2**	**27**	**-19,30%**	**62%**	**-17,70%**	**50%**
	20 minuti	1	1	-29,83%	64%	-50,76%	100%

Persino secondo l'indice di performance in media si può sottolineare un miglioramento significativo per i volumi di ZCash con frequenza a 10 minuti in cui il VAR aggiustato ottiene previsioni migliori in media rispetto al VAR del 19,30% con una robustezza dei risultati del 62%. Di seguito viene mostrato il grafico delle previsioni dei volumi di ZCash con frequenza di osservazioni di 10 minuti in cui si sono individuati miglioramenti per l'indice di performance in media, dei residui e di distanza nella combinazione di previsione-*k* rispettivamente di 2-27.

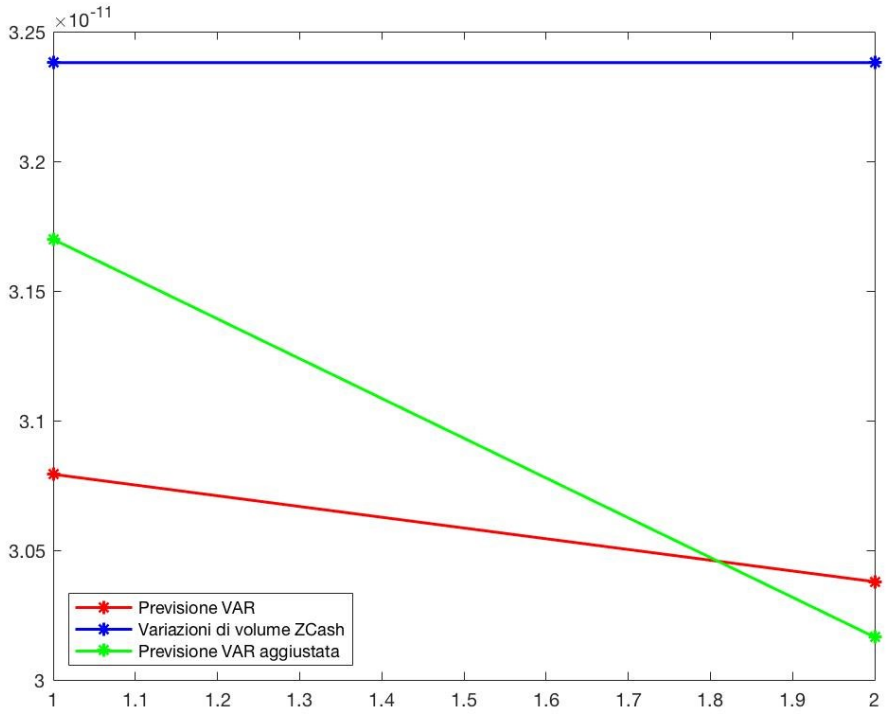

Figura 3.7 - Indice di performance di distanza, volumi V_t ZCash, frequenza osservazioni 10 minuti.

Dalla Figura 3.7 si può notare come la prima previsione aggiustata del VAR (linea verde) sia molto più vicina al valore reale della variazione di volume di ZCash rispetto alla previsione del VAR (linea rossa). Infatti, il miglioramento dell'indice di performance di distanza per tale combinazione di previsione e k è del 50%, perciò la seconda previsione aggiustata risulta più distante dai valori reali rispetto alla previsione effettuata con il VAR. Tuttavia, anche secondo l'indice dei residui le previsioni del VAR aggiustate risultano migliori alle previsioni del VAR del 17,70%. In relazione proprio all'indice di performance dei residui, di seguito vengono riportati nella tabella i risultati dei miglioramenti ottenuti considerando le

combinazioni ottimali di previsioni e *k* per le variazioni di volume delle valute digitali.

Tabella 3.6 - Indice di performance dei residui, volumi V_t, ampiezza previsionale 1 periodo.

		INDICE PERFORMANCE RESIDUI				INDICE PERFORMANCE DISTANZA	INDICE PERFORMANCE MEDIA
	FREQUENZA OSSERVAZIONI	PREVISIONI	K	% MIGLIORAMENTO	ROBUSTEZZA	% MIGLIORAMENTO	% MIGLIORAMENTO
BTC	1 minuto	2	5	-38,86%	22%	50%	-12,93%
	10 minuti	6	45	-0,18%	20%	33%	417,81%
	20 minuti	7	47	-34,54%	20%	71%	-64,25%
ETH	1 minuto	2	33	-21,74%	26%	50%	-57,97%
	10 minuti	4	50	-0,08%	36%	75%	-0,06%
	20 minuti	1	5	-33,63%	42%	100%	-18,53%
LTC	1 minuto	40	47	-0,01%	24%	30%	1,87%
	10 minuti	2	4	-73,16%	28%	50%	-50,64%
	20 minuti	**3**	**49**	**-37,82%**	**24%**	**100%**	**-20,06%**
ZC	1 minuto	31	49	-1,64%	26%	35%	-0,74%
	10 minuti	11	31	-0,31%	44%	64%	109,88%
	20 minuti	4	25	-0,15%	40%	100%	0,05%

Secondo l'indice di performance dei residui, effettuando un'analisi di robustezza non si perviene a dei risultati superiori al 50%. Tuttavia, può essere segnalato il miglioramento riscontrato in Litecoin con frequenza di osservazioni di 20 minuti nella combinazione di previsione-*k* pari a 3-49. Di seguito viene mostrato il grafico delle previsioni del VAR aggiustate con la combinazione sopra citata, per le variazioni di volume di Litecoin con frequenza di osservazione a 20 minuti.

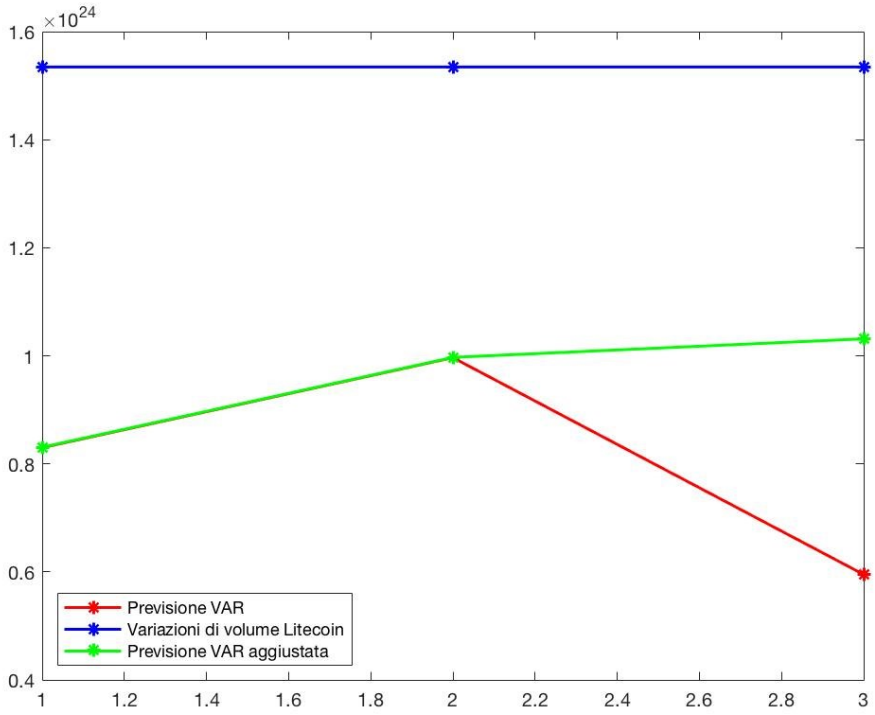

Figura 3.8 - Indice di performance dei residui, volumi V_t Litecoin, frequenza osservazioni 20 minuti.

Come si può notare dalla Figura 3.8, le previsioni aggiustate del VAR (linea verde) hanno distanza minore rispetto alla previsione del VAR (linea rossa). In particolare, il miglioramento si nota nella terza previsione in cui il VAR aggiustato segue il trend delle osservazioni reali dei volumi di Litecoin, mentre il VAR sbaglia in maniera netta come indicato dall' indici di performance in media (-20,06%) e dall'indice di performance dei residui (-37,82%). Tuttavia, il problema del miglioramento viene riscontrato nella sua robustezza (24%) che non raggiunge il valore limite del 50%.

3.2.3 Miglioramenti nei rendimenti: ampiezza previsionale 6 periodi

In questa sezione verranno trattati i miglioramenti ottenuti mediante i tre indici di performance relativamente ai rendimenti R_t delle quattro criptovalute effettuando previsioni con il VAR a 6 periodi. Infatti, come spiegato nel precedente capitolo, il VAR è stato utilizzato per fare previsione scegliendo in questo caso il sesto valore previsto, anche per valutare in modo differente il possibile miglioramento da parte del classificatore KNN.

Tabella 3.7 - Indice di performance di distanza, rendimenti R_t, ampiezza previsionale 6 periodi.

		INDICE PERFORMANCE DISTANZA				INDICE PERFORMANCE RESIDUI	INDICE PERFORMANCE MEDIA
	FREQUENZA OSSERVAZIONI	PREVISIONI	K	% MIGLIORAMENTO	ROBUSTEZZA	% MIGLIORAMENTO	% MIGLIORAMENTO
BTC	1 minuto	6	42	50%	18%	92,87%	-75,33%
	10 minuti	31	40	52%	10%	6,02%	-2,63%
	20 minuti	-	-	-	-	-	-
ETH	1 minuto	50	36	72%	54%	-33,29%	-23,21%
	10 minuti	6	30	83%	42%	-75,75%	-109,99%
	20 minuti	21	8	90%	22%	-47,93%	-158,64%
LTC	**1 minuto**	**37**	**28**	**92%**	**54%**	**-19,26%**	**-12,49%**
	10 minuti	21	36	67%	50%	5,53%	-29,43%
	20 minuti	15	5	53%	2%	28,77%	-23,12%
ZC	1 minuto	48	2	73%	12%	-1,95%	-1,45%
	10 minuti	2	23	50%	38%	96,86%	27,45%
	20 minuti	50	29	86%	26%	-86,48%	-104,70%

Nella Tabella 3.7 sono riportati i risultati ottenuti valutando l'indice di performance di distanza ed ottenendo le migliori combinazioni di previsione-*k* per le quattro criptovalute considerate. In particolare, è da segnalare il miglioramento del 92% ottenuto per Litecoin con frequenza di osservazioni a 1 minuto, di cui in seguito viene mostrato il grafico delle previsioni aggiustate del VAR.

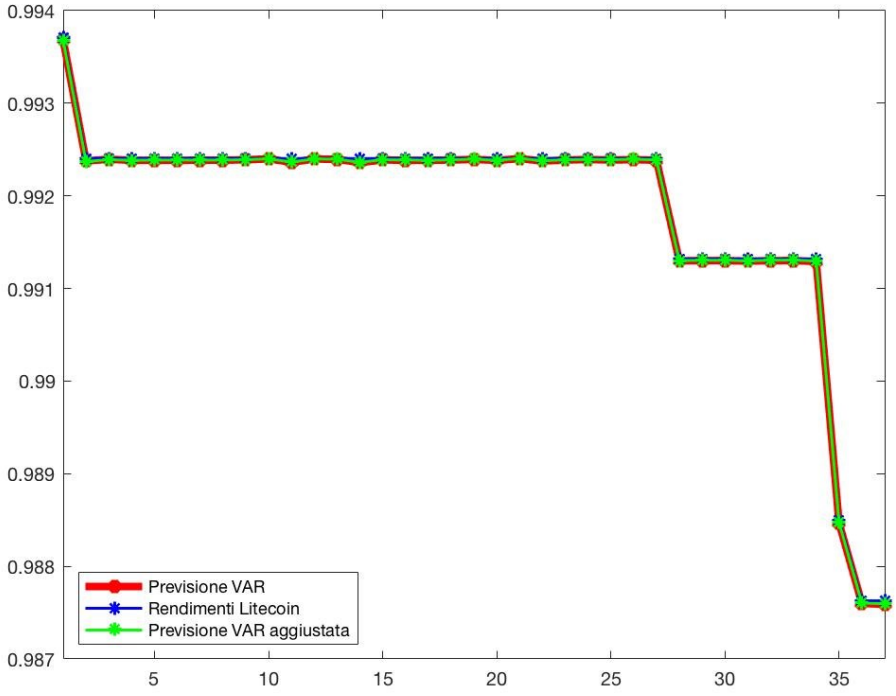

Figura 3.9 - Indice di performance di distanza, rendimenti R_t Litecoin, frequenza osservazioni 1 minuto.

In generale, nella Figura 3.9 le previsioni del VAR sono abbastanza accurate, tuttavia le previsioni aggiustate del VAR risultano essere migliori in media del 12,49% e relativamente alla somma dei quadrati dei residui del 19,26%.

Un altro caso deve essere sottolineato, infatti, nella Tabella 3.7 relativamente a Bitcoin con frequenza di osservazione a 20 minuti, l'indice di performance non ha individuato nessuna miglioria per le 50 previsioni valutate.

Per quanto riguarda l'indice di performance in media, di seguito vengono illustrati i risultati ottenuti per le criptovalute individuando le combinazioni di previsione e *k* ottimali e valutandole con i risultati delle stesse per gli altri indici.

Tabella 3.8 - Indice di performance in media, rendimenti R_t, ampiezza previsionale 6 periodi.

		INDICE PERFORMANCE MEDIA				INDICE PERFORMANCE RESIDUI	INDICE PERFORMANCE DISTANZA
	FREQUENZA OSSERVAZIONI	PREVISIONI	K	% MIGLIORAMENTO	ROBUSTEZZA	% MIGLIORAMENTO	% MIGLIORAMENTO
BTC	1 minuto	49	36	-178,75%	56%	249,03%	16%
	10 minuti	31	40	-2,63%	10%	6,02%	52%
	20 minuti	-	-	-	-	-	-
ETH	**1 minuto**	**50**	**36**	**-23,21%**	**68%**	**-33,29%**	**72%**
	10 minuti	16	26	-333,51%	76%	253,50%	6%
	20 minuti	21	14	-83,41%	56%	-58,52%	67%
LTC	1 minuto	37	9	-55,24%	58%	90,11%	59%
	10 minuti	24	27	-203,81%	58%	-11,48%	63%
	20 minuti	17	6	-129,69%	12%	33,80%	35%
ZC	1 minuto	48	2	-1,45%	4%	-1,95%	73%
	10 minuti	28	21	-7,79%	34%	5,62%	43%
	20 minuti	50	50	-177,09%	76%	-12,94%	66%

Anche nel caso dell'indice di performance in media, in Bitcoin con frequenza di osservazione a 20 minuti non sono stati riscontrati miglioramenti rispetto alle

previsioni effettuate dal VAR. Tuttavia, è da sottolineare il risultato ottenuto per Ethereum con frequenza di osservazioni ad 1 minuto, in cui in media le previsioni aggiustate del VAR sono migliori del 23,21% rispetto alle previsioni effettuate con il VAR. Di seguito viene mostrato il grafico delle previsioni aggiustate dal VAR per Ethereum con frequenza di osservazioni a 1 minuto per la combinazione 50-36 di previsione e *k*.

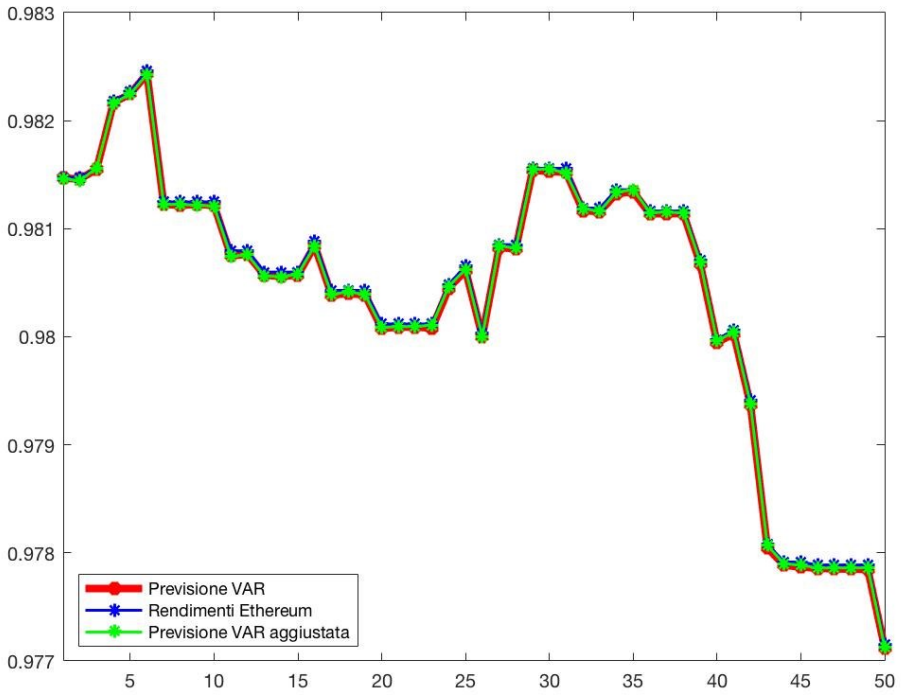

Figura 3.10 - Indice di performance in media, rendimenti R_t Ethereum, frequenza osservazioni 1 minuto.

Come nel caso della Figura 3.9, le previsioni del VAR risultano dal grafico essere accurate, tuttavia secondo i risultati ottenuti, l'indice di performance dei residui indica che le previsioni del VAR sono peggiori del 33,29% e l'indice di distanza

sottolinea che il 72% delle previsioni aggiustate ha distanza minore rispetto alle previsioni del VAR.

Infine, valutando in partenza i miglioramenti dell'indice di performance dei residui per scegliere la combinazione di previsione-*k* ottimale, sono stati ottenuti i risultati mostrati nella seguente tabella.

Tabella 3.9 - Indice di performance dei residui, rendimenti R_t, ampiezza previsionale 6 periodi.

		INDICE PERFORMANCE RESIDUI				INDICE PERFORMANCE DISTANZA	INDICE PERFORMANCE MEDIA
	FREQUENZA OSSERVAZIONI	PREVISIONI	K	% MIGLIORAMENTO	ROBUSTEZZA	% MIGLIORAMENTO	% MIGLIORAMENTO
BTC	1 minuto	11	8	-16,61%	16%	82%	-8,67%
	10 minuti	28	44	-17,83%	10%	61%	-13,31%
	20 minuti	-	-	-	-	-	-
ETH	1 minuto	50	36	-33,29%	38%	72%	-23,21%
	10 minuti	**6**	**30**	**-75,75%**	**40%**	**83%**	**-109,99%**
	20 minuti	21	13	-51,30%	20%	62%	-71,63%
LTC	1 minuto	37	26	-19,26%	44%	92%	-12,49%
	10 minuti	13	30	-55,67%	42%	85%	-97,70%
	20 minuti	20	-	-	-	-	-
ZC	1 minuto	48	2	-1,95%	2%	73%	-1,45%
	10 minuti	26	17	-0,45%	12%	42%	-8,40%
	20 minuti	50	29	-86,48%	26%	6%	-104,70%

Persino nei risultati mostrati nella Tabella 3.9, in Bitcoin con frequenza di osservazione a 20 minuti non sono stati segnalati miglioramenti nell'indice di performance dei residui. Inoltre, in Litecoin con frequenza a 20 minuti, l'analisi di

robustezza per la previsione 20 non ha segnalato alcun miglioramento nelle 50 serie passate, perciò non è stato scelto alcun *k* ottimale.

Tuttavia, di seguito viene mostrato il grafico del caso dei rendimenti di Ethereum con frequenza di osservazione a 10 minuti, in cui sono stati osservati miglioramenti nei tre indici di performance valutati.

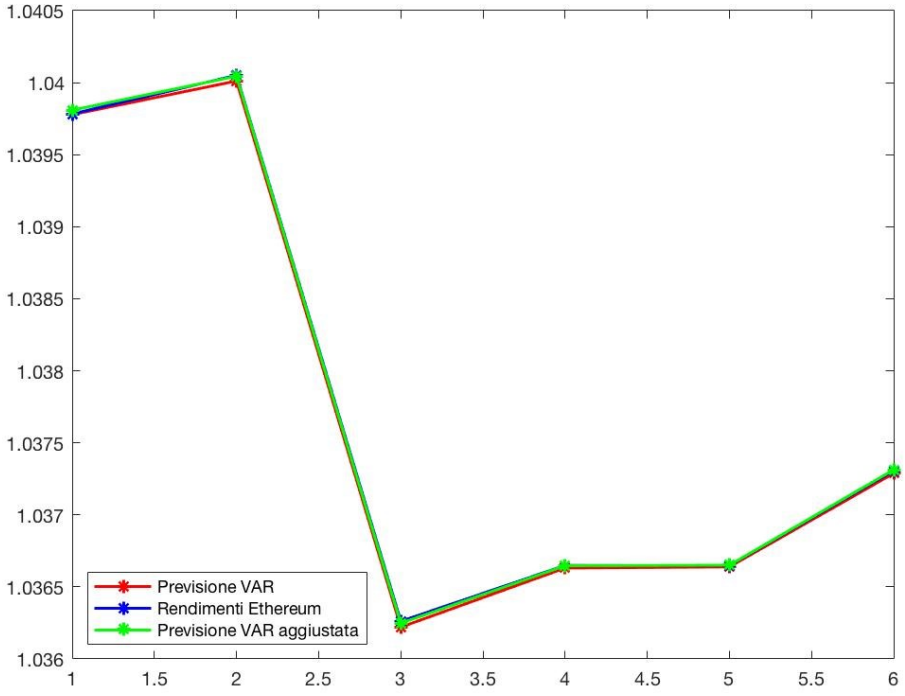

Figura 3.11 - Indice di performance dei residui, rendimenti R_t Ethereum, frequenza osservazioni 10 minuti.

Come si nota dalla Figura 3.11 le previsioni aggiustate del VAR risultano migliori rispetto a quelle effettuate dal VAR. Infatti, secondo l'indice di performance dei residui e in media le previsioni aggiustate sono migliori rispettivamente del 75,75% e del 109,99%, mentre l'82% delle previsioni aggiustate attraverso il KNN risultano

avere distanza minore rispetto alle previsioni del VAR. Tuttavia, in questo caso il problema è situato nella robustezza del risultato, in quanto questa non supera il 50%.

3.2.4 Miglioramenti nei volumi: ampiezza previsionale 6 periodi

Valutando i risultati degli indici di performance sul miglioramento delle variazioni di volume V_t per le quattro criptovalute, utilizzando l'ampiezza previsionale a 6 periodi nelle previsioni effettuate con il VAR, si sono notati maggiori miglioramenti rispetto a quanto descritto per l'ampiezza previsionale ad 1 periodo.

Tabella 3.10 - Indice di performance di distanza, volumi V_t, ampiezza previsionale 6 periodi.

		INDICE PERFORMANCE DISTANZA				INDICE PERFORMANCE RESIDUI	INDICE PERFORMANCE MEDIA
	FREQUENZA OSSERVAZIONI	PREVISIONI	K	% MIGLIORAMENTO	ROBUSTEZZA	% MIGLIORAMENTO	% MIGLIORAMENTO
BTC	1 minuto	2	18	50%	54%	13,79%	-197,82%
	10 minuti	2	20	100%	60%	-42,09%	-58,49%
	20 minuti	6	47	50%	24%	210,11%	-80,87%
ETH	1 minuto	14	7	50%	36%	3,69%	2,37%
	10 minuti	44	11	50%	10%	58,66%	19,66%
	20 minuti	**12**	**41**	**58%**	**60%**	**-31,98%**	**-21,89%**
LTC	1 minuto	14	31	57%	40%	-1,35%	-0,75%
	10 minuti	16	15	50%	28%	-0,03%	1,03%
	20 minuti	4	6	75%	42%	-76,65%	-122,70%
ZC	1 minuto	2	48	50%	74%	-62,16%	-48,52%
	10 minuti	14	43	50%	62%	688,68%	86,90%
	20 minuti	2	8	100%	66%	-94,49%	-78,24%

La Tabella 3.10 mostra i miglioramenti ottenuti valutando l'indice di performance di distanza delle previsioni delle variazioni di volume V_t relative alle quattro criptovalute considerando i valori ottimali di previsione e k. Tra i risultati ottenuti, si può sottolineare come in Ethereum con frequenza 20 minuti per la combinazione di previsione e k pari a 12-41 il miglioramento è stato riscontrato in tutti e tre gli indici di performance considerati, come mostrato nel grafico in seguito.

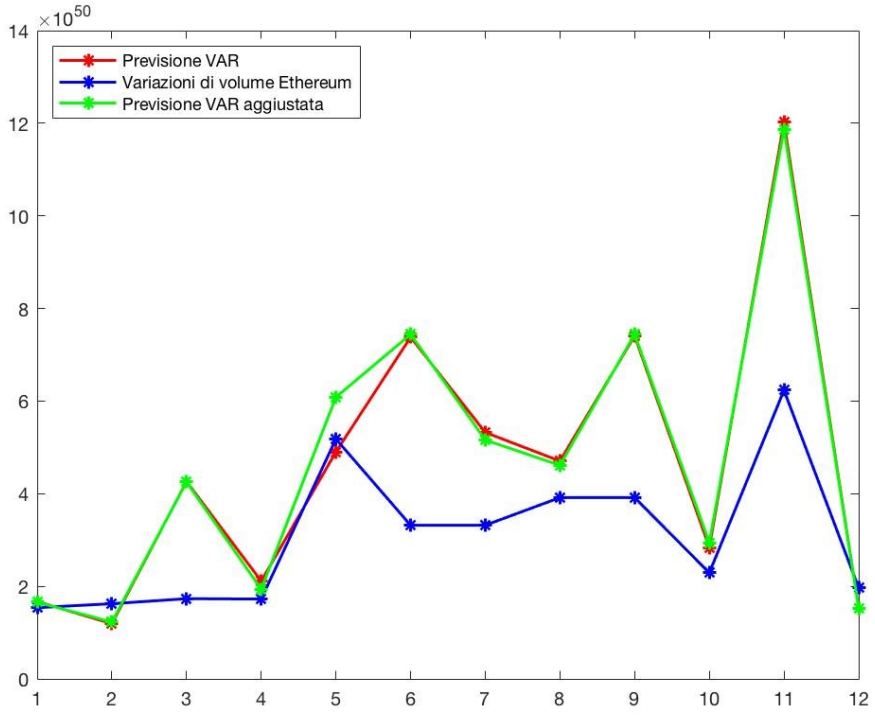

Figura 3.12 - Indice di performance di distanza, volumi V_t Ethereum, frequenza osservazioni 20 minuti.

In questo caso si può notare dalla Figura 3.12 come la previsione del VAR aggiustata abbia in alcuni punti distanza minore dai valori reali delle variazioni di

volume di Ethereum rispetto alle previsioni del VAR con una robustezza del risultato nelle 50 serie precedenti pari al 60%.

Relativamente all'indice di performance in media, di seguito viene riportata la tabella dei risultati per le variazioni di volume delle quattro criptovalute considerate, in cui sono segnalate le combinazioni migliori di previsione e *k*.

Tabella 3.11 - Indice di performance in media, volumi V_t, ampiezza previsionale 6 periodi.

		INDICE PERFORMANCE MEDIA				INDICE PERFORMANCE RESIDUI	INDICE PERFORMANCE DISTANZA
	FREQUENZA OSSERVAZIONI	PREVISIONI	K	% MIGLIORAMENTO	ROBUSTEZZA	% MIGLIORAMENTO	% MIGLIORAMENTO
BTC	1 minuto	2	2	-731,24%	56%	5243,42%	0%
	10 minuti	4	40	-191,51%	26%	34,92%	25%
	20 minuti	7	44	-73,94%	8%	209,82%	43%
ETH	**1 minuto**	**20**	**20**	**-9,91%**	**66%**	**-3,81%**	**50%**
	10 minuti	19	25	-0,02%	24%	0,24%	32%
	20 minuti	46	50	-18,31%	4%	-28,14%	48%
LTC	1 minuto	11	34	-22,95%	60%	2,37%	55%
	10 minuti	1	21	-1119,31%	68%	10289,83%	0%
	20 minuti	40	-	-	-	-	-
ZC	1 minuto	16	41	-14,51%	78%	6,74%	38%
	10 minuti	1	26	-85,59%	60%	-97,92%	100%
	20 minuti	8	45	-26,64%	64%	4,07%	38%

Come si può notare dalla Tabella dei risultati, per Litecoin con frequenza di osservazioni a 20 minuti era stata individuata la previsione ottima (40) con totale miglioramenti massimi, tuttavia, effettuando l'analisi di robustezza non sono stati

riscontrati miglioramenti nelle 50 serie passate, essendo impossibilitati a scegliere un *k* ottimo. Invece, un buon risultato è stato riscontrato in Ethereum con frequenza di osservazioni ad 1 minuto, in cui in media le previsioni aggiustate dal VAR sono migliori del 9,91% rispetto alle previsioni del VAR. Di seguito viene riportato il grafico delle previsioni aggiustate relative alle variazioni dei volumi di Ethereum con frequenza 1 minuto.

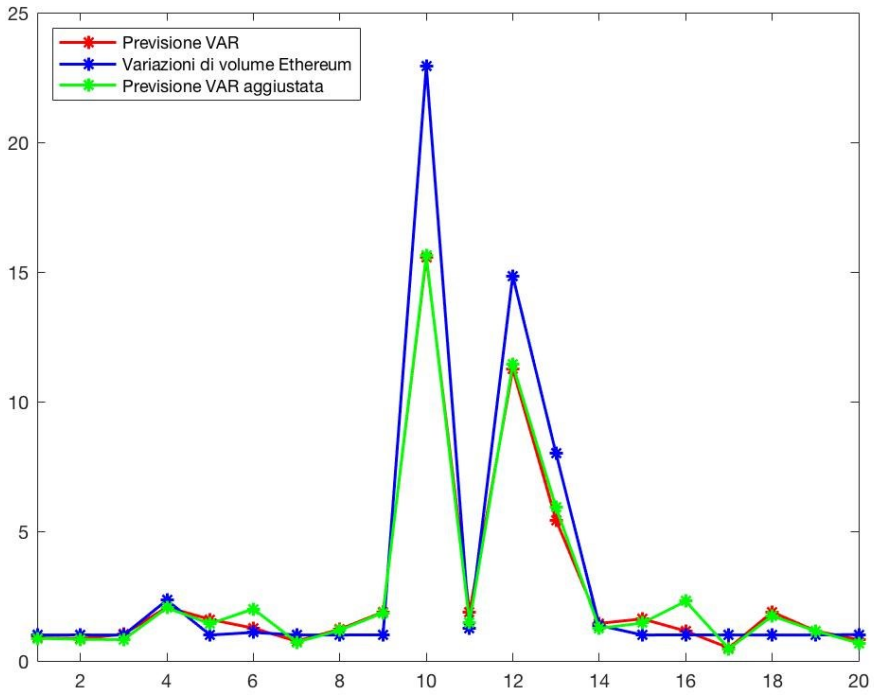

Figura 3.13 - Indice di performance in media, volumi V_t Ethereum, frequenza osservazioni 1 minuto.

Anche in questo caso si segnala che le previsioni aggiustate del VAR sulle variazioni di volume di Ethereum possiedono distanza minore delle previsioni del

VAR nel 50% dei casi con una robustezza dei risultati riguardante l'indice di performance in media del 66%.

Per ultimo, vengono illustrati i risultati delle previsioni e del valore di *k* scelti andando a valutare l'indice di performance dei residui e confrontando quest'ultimi con i risultati degli altri due indici per le stesse combinazioni individuate.

Tabella 3.12 - Indice di performance dei residui, volumi V_t, ampiezza previsionale 6 periodi.

		INDICE PERFORMANCE RESIDUI				INDICE PERFORMANCE DISTANZA	INDICE PERFORMANCE MEDIA
	FREQUENZA OSSERVAZIONI	PREVISIONI	K	% MIGLIORAMENTO	ROBUSTEZZA	% MIGLIORAMENTO	% MIGLIORAMENTO
BTC	1 minuto	1	2	-90,97%	34%	100%	-69,96%
	10 minuti	2	20	-42,09%	30%	100%	-58,49%
	20 minuti	1	50	-85,43%	28%	100%	-138,17%
ETH	1 minuto	2	48	-81,53%	42%	100%	189,75%
	10 minuti	16	25	-0,03%	6%	38%	-0,36%
	20 minuti	46	46	-21,47%	26%	35%	-15,19%
LTC	1 minuto	14	21	-1,35%	32%	57%	-0,75%
	10 minuti	16	9	-0,03%	12%	50%	1,03%
	20 minuti	1	47	0,00%	44%	100%	0,00%
ZC	1 minuto	1	48	-64,05%	50%	100%	-40,04%
	10 minuti	1	9	-1,45%	50%	100%	-0,73%
	20 minuti	**3**	**9**	**-55,24%**	**34%**	**67%**	**-120,45%**

Dai risultati ottenuti si possono segnalare le combinazioni di previsione e *k* scelte per ZCash con frequenza di osservazione a 1 minuto e a 10 minuti, in cui l'indice di performance dei residui indica che le previsioni del VAR aggiustato sono

migliori rispetto a quelle effettuate mediante il VAR, con robustezza dei risultati al 50%.

Inoltre, di seguito viene mostrato il grafico relativo alle variazioni di volume di ZCash con frequenza di osservazione a 20 minuti, in cui per la previsione a 3 periodi e con *k* pari a 9 sono stati ottenuti miglioramenti significativi per tutti e tre gli indici di performance, malgrado la robustezza dell'indice dei residui sia solo del 34%.

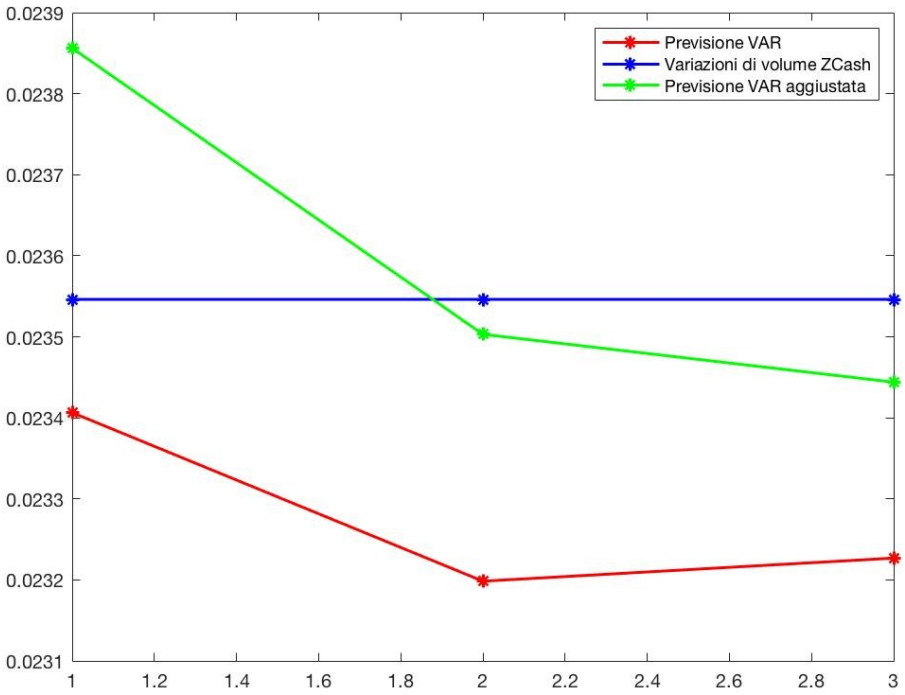

Figura 3.14 - Indice di performance dei residui, volumi V_t ZCash, frequenza osservazioni 20 minuti.

Come si può notare dalla Figura 3.14, la previsione aggiustata del VAR risulta migliore nelle ultime due previsioni per distanza dai valori reali rispetto alla previsione del VAR risultando quest'ultima peggiore in media del 120,45%.

CONCLUSIONI

Tra il 2017 e il 2018, il mercato delle criptovalute, data l'ampia dimensione dei suoi guadagni e delle sue perdite ha catturato l'attenzione degli investitori.

Proprio per la notevole importanza che hanno assunto le criptovalute, è stata ricercata una possibile relazione tra il mercato delle stesse e il social media Twitter. Utilizzando un approccio basato sul lessico per convertire i tweet in punteggi combinato ad un modello vettoriale autoregressivo VAR bivariato, sono state individuate molteplici relazioni di causa-effetto tra le variabili delle criptovalute considerate. In particolare, in seguito alle analisi svolte mediante test di Granger è stato individuato come le variabili derivanti da Twitter, score e volume dei tweet, causino i rendimenti e le variazioni di volume delle criptovalute. In questo senso, Twitter acquisisce un importante potere predittivo circa il mercato delle criptovalute. Tuttavia, dalle analisi effettuate si è anche evidenziato come le informazioni derivanti dal social media Twitter siano guidate dal mercato delle criptovalute, in quanto anche le variabili finanziarie "granger causano" gli score e il volume dei tweet. In generale, i risultati migliori ottenuti dimostrano che le relazioni di causa effetto tra il mercato delle criptovalute e Twitter si attestano nelle variazioni di volume delle valute digitali. Questo può avere un'importanza rilevante

per gli investitori che hanno l'intento di individuare dei possibili andamenti futuri, in quanto un alto volume di tweet emessi può comportare un elevato volume di criptovalute scambiate con conseguente oscillazione dei prezzi delle stesse.

Invece, relativamente ai miglioramenti delle previsioni ottenuti tramite K-Nearest-Neighbor, sono stati segnalati alcuni risultati positivi per le criptovalute considerate. Infatti, andando a studiare diverse frequenze di osservazioni dei dati e differenti ampiezze previsionali adottate nel metodo parametrico, è stato possibile evidenziare come in particolare si siano riscontrati miglioramenti per le variazioni di volume delle criptovalute. In aggiunta, sono stati sottolineati maggiori miglioramenti in seguito all'applicazione del metodo KNN nel caso in cui attraverso il VAR siano state effettuate previsioni a 6 periodi.

Queste considerazioni finali possono essere di notevole importanza, insieme alla capacità predittiva di Twitter sulle criptovalute, per supportare gli investitori nelle strategie di trading.

Pertanto, il vantaggio di analizzare l'andamento delle criptovalute con frequenza di osservazioni al minuto è quello di valutare al meglio il loro comportamento.

Tuttavia, la disponibilità e la gestione dei dati necessari per effettuare tali analisi risulta essere un fattore determinante in quanto, se si intendesse espandere l'ampiezza delle osservazioni analizzate, si dovrebbe aumentare la capacità di gestire una quantità di dati considerevole.

APPENDICE

Tabella A.1 - P-value test di Granger, Bitcoin frequenza osservazioni 1 minuto.

RELAZIONI	P-VALUE														
	ORDINI DEL VAR (p)														
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
$S_t \rightarrow R_t$	0,780	0,619	0,664	0,802	0,889	0,554	0,376	0,470	0,566	0,512	0,544	0,615	0,687	0,741	0,550
$R_t \rightarrow S_t$	0,193	0,309	0,374	0,392	0,286	0,376	0,417	0,408	0,503	0,585	0,687	0,758	0,785	0,830	0,821
$TV_t \rightarrow V_t$	0,234	0,329	0,493	0,524	0,726	0,819	0,732	0,768	0,814	0,867	0,664	0,733	0,699	0,667	0,485
$V_t \rightarrow TV_t$	0,549	0,417	0,685	0,609	0,611	0,626	0,516	0,540	0,527	0,627	0,702	0,770	0,839	0,893	0,761
$S_t \rightarrow V_t$	0,595	0,768	0,561	0,516	0,490	0,139	0,164	0,136	0,131	0,208	0,260	0,137	0,187	0,218	0,261
$V_t \rightarrow S_t$	0,853	0,867	0,983	0,681	0,449	0,502	0,567	0,203	0,238	0,355	0,409	0,347	0,414	0,415	0,481
$TV_t \rightarrow R_t$	0,427	0,702	0,856	0,187	0,203	0,250	0,144	0,078	0,088	0,076	0,050	**0,025**	**0,033**	0,045	**0,028**
$R_t \rightarrow TV_t$	0,739	0,403	0,245	0,338	0,436	0,573	0,510	0,596	0,763	0,760	0,734	0,660	0,734	0,839	0,782

Tabella A.2 - P-value test di Granger, Bitcoin frequenza osservazioni 10 minuti.

RELAZIONI	P-VALUE														
	ORDINI DEL VAR (p)														
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
$S_t \rightarrow R_t$	0,259	0,488	0,682	0,824	0,609	0,591	0,681	0,536	0,628	0,720	0,775	0,821	0,821	0,872	0,906
$R_t \rightarrow S_t$	0,547	0,595	0,373	0,446	0,586	0,597	0,569	0,699	0,708	0,724	0,798	0,502	0,563	0,523	0,503
$TV_t \rightarrow V_t$	0,105	0,170	0,138	0,067	0,054	0,098	0,104	0,120	0,178	0,125	0,173	0,197	0,214	0,160	0,214
$V_t \rightarrow TV_t$	0,129	0,134	0,209	0,380	0,389	0,377	0,511	0,476	0,564	0,582	0,583	0,604	0,441	0,377	0,448
$S_t \rightarrow V_t$	0,282	**0,049**	0,099	**0,027**	**0,034**	**0,036**	0,058	0,057	**0,040**	0,053	0,088	0,125	0,166	0,210	0,276
$V_t \rightarrow S_t$	0,703	0,254	0,135	0,218	0,287	0,339	0,419	0,563	0,674	0,692	0,569	0,620	0,687	0,453	0,488
$TV_t \rightarrow R_t$	0,055	0,175	0,162	0,184	0,277	0,361	0,448	0,549	0,374	0,411	0,448	0,495	0,584	0,605	0,534
$R_t \rightarrow TV_t$	0,786	0,239	0,371	0,515	0,557	0,668	0,750	0,815	0,852	0,864	0,799	0,565	0,601	0,630	0,595

Tabella A.3 - P-value test di Granger, Bitcoin frequenza osservazioni 20 minuti.

RELAZIONI	P-VALUE														
	ORDINI DEL VAR (p)														
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
$S_t \rightarrow R_t$	0,529	0,810	0,769	0,857	0,932	0,957	0,978	0,991	0,924	0,957	0,948	0,971	0,982	0,922	0,893
$R_t \rightarrow S_t$	0,382	0,633	0,812	0,861	0,869	0,658	0,490	0,274	0,296	0,363	0,437	0,442	0,524	0,451	0,428
$TV_t \rightarrow V_t$	0,777	0,619	0,578	0,761	0,402	0,406	0,247	0,335	0,410	0,409	0,374	0,435	0,165	0,213	0,307
$V_t \rightarrow TV_t$	0,066	0,113	0,121	0,159	0,183	0,398	0,509	0,420	0,574	0,685	0,493	0,361	0,369	0,413	0,575
$S_t \rightarrow V_t$	0,365	0,548	0,634	0,772	0,786	0,841	0,889	0,918	0,686	0,653	0,488	0,475	0,548	0,621	0,688
$V_t \rightarrow S_t$	0,079	0,139	0,229	0,209	0,323	0,363	0,479	0,551	0,595	0,621	0,685	0,745	0,805	0,855	0,546
$TV_t \rightarrow R_t$	0,379	0,400	0,566	0,701	0,744	0,824	0,862	0,912	0,776	0,879	0,894	0,869	0,731	0,696	0,740
$R_t \rightarrow TV_t$	0,419	0,851	0,944	0,883	0,909	0,585	0,700	0,584	0,635	0,667	0,542	0,600	0,731	0,804	0,881

Tabella A.4 - P-value test di Granger, Ethereum frequenza osservazioni 1 minuto.

RELAZIONI	P-VALUE														
	ORDINI DEL VAR (p)														
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
$S_t \rightarrow R_t$	0,668	0,590	0,807	0,908	0,961	0,830	0,882	0,665	0,671	0,756	0,754	0,809	0,843	0,878	0,904
$R_t \rightarrow S_t$	0,053	**0,039**	0,108	0,155	0,105	0,118	0,185	0,246	0,351	0,440	0,471	0,548	0,672	0,545	0,626
$TV_t \rightarrow V_t$	0,997	0,870	0,414	0,428	0,594	0,712	0,820	0,812	0,476	0,514	0,611	0,574	0,500	0,516	0,592
$V_t \rightarrow TV_t$	0,128	0,196	0,382	0,318	0,537	0,658	0,654	0,702	0,712	0,761	0,829	0,830	0,883	0,391	0,461
$S_t \rightarrow V_t$	0,962	0,994	0,786	0,911	0,980	0,828	0,687	0,769	0,770	0,270	0,148	0,103	0,138	0,179	0,213
$V_t \rightarrow S_t$	0,868	0,772	0,869	0,940	0,917	0,529	0,642	0,614	0,478	0,458	0,564	0,595	0,692	0,769	0,580
$TV_t \rightarrow R_t$	0,884	0,981	0,974	0,989	0,564	0,245	0,346	0,332	0,350	0,396	0,415	0,411	0,317	0,229	0,282
$R_t \rightarrow TV_t$	0,283	0,253	0,442	0,448	0,578	0,555	0,754	0,780	0,483	0,266	0,188	0,245	0,256	0,252	0,256

Tabella A.5 - P-value test di Granger, Ethereum frequenza osservazioni 10 minuti.

RELAZIONI	P-VALUE														
	ORDINI DEL VAR (p)														
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
$S_t \rightarrow R_t$	0,462	0,580	0,417	0,471	0,548	0,805	0,919	0,960	0,929	0,905	0,700	0,757	0,691	0,221	0,214
$R_t \rightarrow S_t$	0,260	0,262	0,172	0,255	0,330	0,158	0,183	0,316	0,397	0,234	0,273	0,168	0,222	0,224	0,259
$TV_t \rightarrow V_t$	0,736	0,317	**0,012**	**0,019**	**0,043**	**0,028**	**0,029**	**0,044**	**0,030**	0,057	0,081	0,126	0,072	0,059	0,073
$V_t \rightarrow TV_t$	0,646	0,207	0,345	0,239	0,287	0,144	0,156	0,264	0,176	0,237	0,274	0,383	0,454	0,537	0,592
$S_t \rightarrow V_t$	0,720	0,863	0,689	0,820	0,637	0,718	0,644	0,753	0,708	0,784	0,851	0,870	0,720	0,441	0,463
$V_t \rightarrow S_t$	0,794	0,875	0,963	0,979	0,984	0,990	0,995	0,988	0,991	0,950	0,973	0,970	0,977	0,985	0,986
$TV_t \rightarrow R_t$	0,274	0,530	0,675	0,715	0,547	0,669	0,761	0,819	0,854	0,859	0,831	0,877	0,928	0,937	0,690
$R_t \rightarrow TV_t$	0,788	0,752	0,248	0,316	0,411	0,310	0,387	0,435	0,345	0,331	0,321	0,264	0,373	0,206	0,283

Tabella A.6 - P-value test di Granger, Ethereum frequenza osservazioni 20 minuti.

RELAZIONI	P-VALUE														
	ORDINI DEL VAR (p)														
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
$S_t \rightarrow R_t$	0,524	0,725	0,867	0,953	0,828	0,955	0,278	0,369	0,430	0,488	0,566	0,595	0,545	0,645	0,662
$R_t \rightarrow S_t$	0,590	0,577	0,246	0,347	0,231	0,269	0,364	0,461	0,659	0,616	0,735	0,813	0,755	0,815	0,721
$TV_t \rightarrow V_t$	**0,024**	**0,028**	0,158	0,233	0,235	0,215	0,122	0,161	0,134	0,236	0,298	0,362	0,500	0,404	0,386
$V_t \rightarrow TV_t$	0,064	**0,031**	0,058	0,140	0,086	0,218	0,233	0,188	0,329	0,412	0,316	0,420	0,403	0,398	0,403
$S_t \rightarrow V_t$	0,158	0,239	0,299	0,385	0,442	0,601	0,526	0,478	0,585	0,646	0,707	0,766	0,828	0,564	0,545
$V_t \rightarrow S_t$	0,919	0,996	0,981	0,267	0,150	0,185	0,235	0,142	0,226	0,230	0,193	0,256	0,313	0,192	0,255
$TV_t \rightarrow R_t$	0,553	0,559	0,828	0,945	0,958	0,958	0,848	0,887	0,928	0,951	0,214	0,271	0,337	0,335	0,339
$R_t \rightarrow TV_t$	0,273	0,608	0,359	0,489	0,336	0,417	0,312	0,303	0,264	0,209	0,134	0,187	0,299	0,367	0,253

Tabella A.7 - P-value test di Granger, Litecoin frequenza osservazioni 1 minuto.

RELAZIONI	P-VALUE														
	ORDINI DEL VAR (p)														
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
$S_t \rightarrow R_t$	0,671	0,318	0,509	0,330	0,432	0,352	0,166	0,201	0,256	0,168	0,227	0,289	0,363	0,444	0,450
$R_t \rightarrow S_t$	0,454	0,598	0,821	0,942	0,899	0,789	0,814	0,669	0,401	0,407	0,483	0,589	0,664	0,684	0,760
$TV_t \rightarrow V_t$	0,573	0,249	0,116	0,142	0,247	0,309	0,241	0,179	0,215	0,302	0,079	0,115	**0,032**	**0,038**	0,052
$V_t \rightarrow TV_t$	0,070	0,211	0,144	0,288	0,483	0,453	0,541	0,600	0,513	0,630	0,675	0,733	0,695	0,684	0,563
$S_t \rightarrow V_t$	0,106	0,202	0,327	0,595	0,344	0,139	0,069	0,134	0,138	**0,037**	**0,022**	**0,032**	0,050	0,076	0,097
$V_t \rightarrow S_t$	0,241	0,818	0,216	0,368	0,395	0,598	0,464	0,596	0,542	0,515	0,586	0,649	0,731	0,697	0,677
$TV_t \rightarrow R_t$	0,492	0,535	0,645	0,526	0,671	0,236	0,098	0,143	0,202	0,155	0,141	0,076	0,107	0,145	0,177
$R_t \rightarrow TV_t$	0,644	0,661	0,519	0,583	0,721	0,817	0,883	0,957	0,823	0,457	0,530	0,551	0,605	0,663	0,652

Tabella A.8 - P-value test di Granger, Litecoin frequenza osservazioni 10 minuti.

RELAZIONI	P-VALUE														
	ORDINI DEL VAR (p)														
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
$S_t \rightarrow R_t$	**0,043**	0,135	0,273	0,428	0,527	0,490	0,518	0,593	0,615	0,677	0,674	0,697	0,734	0,714	0,614
$R_t \rightarrow S_t$	0,163	0,251	0,402	0,111	0,228	0,275	0,347	0,400	0,486	0,557	0,454	0,422	0,474	0,479	0,093
$TV_t \rightarrow V_t$	0,671	0,855	0,941	0,980	0,611	0,094	0,114	0,162	0,212	0,254	0,297	0,235	0,272	0,333	0,234
$V_t \rightarrow TV_t$	0,507	0,572	0,524	0,670	0,694	0,727	0,734	0,742	0,842	0,889	0,839	0,605	0,420	0,473	0,351
$S_t \rightarrow V_t$	0,568	0,747	0,878	0,874	0,945	0,838	0,828	0,893	0,634	0,700	0,553	0,325	0,308	0,281	0,302
$V_t \rightarrow S_t$	0,763	0,756	0,695	0,579	0,601	0,467	0,448	0,546	0,621	0,512	0,570	0,453	0,552	0,608	0,373
$TV_t \rightarrow R_t$	0,669	0,756	0,867	0,944	0,920	0,762	0,828	0,810	0,842	0,890	0,930	0,943	0,961	0,976	0,957
$R_t \rightarrow TV_t$	0,304	0,500	0,678	0,648	0,770	0,726	0,343	0,435	0,441	0,570	0,502	0,572	0,658	0,714	0,736

Tabella A.9 - P-value test di Granger, Litecoin frequenza osservazioni 20 minuti.

RELAZIONI	P-VALUE														
	ORDINI DEL VAR (p)														
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
$S_t \rightarrow R_t$	0,866	0,101	0,190	0,260	0,345	0,532	0,585	0,699	0,638	0,642	0,597	0,668	0,597	0,556	0,618
$R_t \rightarrow S_t$	0,325	0,377	0,582	0,119	0,165	0,265	0,169	0,168	0,147	0,172	0,196	0,254	0,268	0,202	0,181
$TV_t \rightarrow V_t$	0,403	0,252	0,236	0,395	0,370	0,352	0,311	0,178	0,214	0,282	0,316	0,311	0,398	0,224	0,302
$V_t \rightarrow TV_t$	0,840	0,492	0,776	0,789	0,843	0,836	0,677	0,779	0,874	0,883	0,299	0,149	0,266	0,331	0,352
$S_t \rightarrow V_t$	0,186	0,419	0,596	0,723	0,735	0,787	0,494	0,598	0,482	0,621	0,446	0,501	0,525	0,412	0,341
$V_t \rightarrow S_t$	0,696	**0,039**	**0,039**	**0,046**	0,086	0,134	0,151	0,177	0,237	0,308	0,308	0,368	0,426	0,505	0,531
$TV_t \rightarrow R_t$	0,384	0,464	0,595	0,420	0,394	0,504	0,328	0,420	0,386	0,370	0,468	0,514	0,377	0,418	0,460
$R_t \rightarrow TV_t$	0,402	0,518	0,147	0,113	0,149	0,183	**0,042**	0,082	0,138	0,193	0,194	0,222	0,270	0,164	0,122

Tabella A.10 - P-value test di Granger, ZCash frequenza osservazioni 1 minuto.

RELAZIONI	P-VALUE														
	ORDINI DEL VAR (p)														
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
$S_t \rightarrow R_t$	0,432	0,720	0,442	0,449	0,481	0,620	0,707	0,767	0,716	0,548	0,621	0,514	0,397	0,465	0,403
$R_t \rightarrow S_t$	0,872	0,736	0,140	0,223	0,319	0,492	0,232	0,329	0,396	0,455	0,487	0,424	0,477	0,569	0,594
$TV_t \rightarrow V_t$	0,474	0,697	0,726	0,868	0,929	0,729	0,750	0,801	0,869	0,824	0,621	0,721	0,470	0,553	0,609
$V_t \rightarrow TV_t$	0,254	0,533	0,618	0,804	0,883	0,920	0,937	0,965	0,976	0,889	0,891	0,867	0,882	0,881	0,909
$S_t \rightarrow V_t$	0,207	0,809	0,358	0,569	0,586	0,269	0,351	0,365	0,621	0,425	0,503	0,564	0,341	0,291	0,314
$V_t \rightarrow S_t$	**0,027**	0,114	0,164	0,285	0,146	0,172	0,133	0,191	0,344	0,354	0,418	0,145	0,232	0,227	0,231
$TV_t \rightarrow R_t$	0,743	0,754	0,832	0,929	0,932	0,753	0,840	0,830	0,889	0,901	0,930	0,919	0,950	0,932	0,956
$R_t \rightarrow TV_t$	0,088	0,096	0,107	0,205	0,253	0,420	0,207	0,229	0,309	0,359	0,415	0,467	0,535	0,610	0,692

Tabella A.11 - P-value test di Granger, ZCash frequenza osservazioni 10 minuti.

RELAZIONI	P-VALUE														
	ORDINI DEL VAR (p)														
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
$S_t \rightarrow R_t$	0,742	0,850	0,781	0,629	0,679	0,724	0,769	0,783	0,746	0,813	0,821	0,692	0,588	0,588	0,659
$R_t \rightarrow S_t$	0,948	0,713	0,846	0,728	0,746	0,817	0,786	0,731	0,778	0,833	0,890	0,931	0,894	0,912	0,852
$TV_t \rightarrow V_t$	0,226	0,367	0,091	0,124	0,195	0,280	0,322	0,397	0,418	0,507	0,545	0,210	0,295	0,188	0,149
$V_t \rightarrow TV_t$	**0,008**	**0,016**	**0,039**	0,069	0,111	0,159	0,205	0,265	0,060	0,062	0,076	0,108	0,150	0,120	**0,050**
$S_t \rightarrow V_t$	0,432	0,646	0,072	0,111	**0,038**	**0,009**	**0,017**	**0,026**	**0,041**	0,064	0,087	**0,012**	**0,001**	**0,001**	**0,001**
$V_t \rightarrow S_t$	0,081	0,206	0,324	0,413	0,437	0,592	0,697	0,462	0,456	0,437	0,476	0,564	0,556	0,579	0,611
$TV_t \rightarrow R_t$	0,332	0,407	0,539	0,702	0,702	0,702	0,764	0,857	0,907	0,939	0,952	0,959	0,974	0,982	0,967
$R_t \rightarrow TV_t$	0,377	0,467	0,676	0,800	0,712	0,677	0,479	0,590	0,687	0,715	0,766	0,657	0,622	0,170	0,195

Tabella A.12 - P-value test di Granger, ZCash frequenza osservazioni 20 minuti.

RELAZIONI	P-VALUE														
	ORDINI DEL VAR (p)														
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
$S_t \rightarrow R_t$	0,769	0,419	0,523	0,673	0,799	0,881	0,682	0,713	0,796	0,841	0,877	0,920	0,934	0,936	0,955
$R_t \rightarrow S_t$	0,106	0,188	0,282	0,364	0,508	0,459	0,312	0,368	0,463	0,453	0,515	0,416	0,312	0,347	0,417
$TV_t \rightarrow V_t$	0,833	0,951	0,992	0,995	0,998	0,991	0,826	0,722	0,656	0,784	0,790	0,823	0,859	0,814	0,850
$V_t \rightarrow TV_t$	0,058	0,147	0,218	0,219	0,302	0,359	0,098	0,138	0,224	0,106	0,137	0,180	0,081	**0,045**	0,059
$S_t \rightarrow V_t$	0,742	0,281	0,293	0,355	0,409	0,210	0,309	0,375	0,471	0,515	0,524	0,450	0,523	0,234	0,266
$V_t \rightarrow S_t$	0,384	0,598	0,800	0,665	0,792	0,740	0,791	0,812	0,761	0,774	0,811	0,684	0,774	0,810	0,876
$TV_t \rightarrow R_t$	0,496	0,793	0,779	0,856	0,941	0,947	0,949	0,980	0,984	0,988	0,817	0,861	0,901	0,872	0,904
$R_t \rightarrow TV_t$	0,642	0,736	0,773	0,749	0,858	0,801	0,172	0,234	0,241	0,325	0,331	0,350	0,247	0,257	0,282

BIBLIOGRAFIA

C. Brooks, *"Introductory Econometrics for Finance"*, Second Edition, Cambridge, 2008.

J. Hamilton, *"Time Series Analysis"*, Princeton, 1994.

H. Lütkepohl, *"New Introduction to Multiple Time Series Analysis"*, Springer, 2005.

R. Lucchetti, *"Appunti di analisi delle serie storiche"*, 2015.

C. W. J. Granger, *"Investigating Causal Relations by Econometric Models and Cross-spectral Methods"*, Econometrica, Vol. 37, No. 3, pp. 424-438, 1969.

C. A. Sims, *"Macreoconomic and Reality"*, Econometrica, Vol.48, No. 1, pp. 1-48, 1980.

G. Gallo, B. Pacini, *"Metodi quantitativi per i mercati finanziari"*, Carocci Editore, 2015.

M. Ooms, *"Empirical Vector Autoregressive Modeling"*, Springer, 1994.

P. Brockwell, R. Davis, *"Time Series: Theory and Methods"*, Second Edition, Springer, 1991.

T. M. Mitchell, *"Machine Learning"*, McGraw-Hill, 1997.

J. Han, M. Kamber, J. Pei, *"Data Mining Concept and Techniques"*, Third Edition, Morgan Kaufman, 2012.

G. James, D. Witten, T. Hastie, R. Tibshirani, *"An Introduction to Statistical Learning"*, Springer, 2013.

O. Kraaijeveld, J. De Smedt, *"The predictive power of public Twitter sentiment for forecasting cryptocurrency prices"*, 2020.

S. Nakamoto, *"Bitcoin: A Peer-to-Peer Electronic Cash System"*, 2008.

A. Antonopoulos, *"Mastering Bitcoin: Programming the Open Blockchain"*, O'Reily, Second Edition, 2017.

G. Giudici, A. Milne, D. Vinogradov, *"Cryptocurrencies: market analysis and perspectives"*, 2019.

V. A. Kharde, S. S. Sonowane, *"Sentiment Analysis of Twitter Data: A Survey of Techniques"*, 2016.

V. Pagolu, K. Challa, G. Panda, B. Majhi, *"Sentiment Analysis of Twitter Data for Predicting Stock Market Movements"*, 2016.

A. Agarwal, R. Passonneau, O. Rambow, *"Sentiment Analysis of Twitter Data"*, 2011.

P. Ciaian, M. Rajcaniova & d'Artis Kancs, *"The economics of Bitcoin price formation"*, 2016.

C. Hutto, E. Gilbert, *"VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text"*, 2015.

SITOGRAFIA

https://www.cryptodatadownload.com/index.html

https://it.finance.yahoo.com/

https://developer.twitter.com/en

https://bitinfocharts.com/

https://bitcoin.org/it/

https://coinmarketcap.com/

https://it.mathworks.com/products/datafeed.html

https://it.mathworks.com/products/text-analytics.html