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

Statistica per l'Economia e l'Impresa

RELAZIONE FINALE

INDICI DI BILANCIO, EQUITY SCREENING E STOCK PICKING: UN'APPLICAZIONE AD AZIONI AMERICANE

Relatore Prof. Massimiliano Caporin
Dipartimento di Scienze Statistiche

Laureando: Alessandro Magnabosco
Matricola N° 1217945

Anno Accademico 2021/2022

Indice

INTRODUZIONE

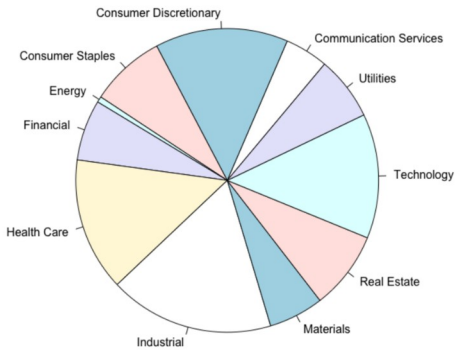
Il seguente lavoro è nato dalla curiosità e dall'interesse riguardante una possibile risoluzione di uno dei problemi più comuni quando ci si approccia all'ambito degli investimenti: come selezionare un gruppo di aziende su cui investire a partire da un paniere più ampio.
L'idea si basa sulla volontà di classificare le aziende quotate sul mercato azionario americano in funzione di alcuni indici di bilancio ritenuti significativi, con l'obiettivo di capire se esse siano classificabili in sottogruppi e, eventualmente, di vedere se tra questi gruppi ce ne siano alcuni comprendenti azioni "migliori" e altri comprendenti azioni "peggiori".
Per la classificazione delle aziende in sottogruppi si è svolta un'analisi di *clustering* utilizzando l'algoritmo *k-means*.
Le variabili sulle quali si è applicata l'analisi di *clustering* sono per la maggior parte indici di bilancio, scelti con l'obiettivo di permettere la valutazione delle singole azioni dal punto di vista del loro prezzaggio, della redditività, della salute aziendale, della possibilità di crescita e della distribuzione di dividendi. Altri indicatori sono stati costruiti con l'obiettivo di valutare il rischio e il rendimento dei singoli titoli.
Sui titoli appartenenti all'indice *Standard & Poor's 500* si è svolta inizialmente un'analisi di *clustering* preliminare su due istanti temporali notevoli, e successivamente una seconda analisi di *clustering* avente come obiettivo la classificazione delle aziende e l'analisi dei singoli gruppi per cercare di capire se vi siano differenze qualitative tra essi.
In conclusione, con i gruppi di titoli considerati migliori, si sono costruiti dei portafogli di investimento e si sono confrontate le performance di questi ultimi con quelle di alcuni portafogli utilizzati come metro di paragone.

1 DESCRIZIONE DEL DATASET

1.1 Descrizione dei dati

I dati utilizzati per svolgere le analisi sono stati scaricati tramite il software *Refinitiv Eikon*, grazie al quale è stato possibile disporre di tutte le quantità necessarie alla redazione di questa tesi di laurea.
Si sono scelti come target dell'analisi tutte le aziende appartenenti all'indice *Standard & Poor's 500*[1].
L'S&P 500 è un indice del mercato azionario americano all'interno del quale rientrano le 500 aziende a più alta capitalizzazione flottante tra quelle quotate a New York. Le aziende appartenenti a questo indice rappresentano circa l'80% del mercato azionario americano, rendendo l'S&P500 un indice piuttosto rappresentativo del mercato statunitense.
Per tutte le aziende sono stati scaricati degli indici di bilancio in serie storica a frequenza mensile a partire da dicembre 2015 fino a dicembre 2021, mentre per quanto riguarda i prezzi dei titoli azionari relativi alle suddette aziende, si è deciso di scaricarli in serie storica a frequenza giornaliera a partire dal 31 dicembre 2014 fino al 31 dicembre 2021.
Inoltre, ad ogni azienda è stato associato il settore economico all'interno del quale essa opera. In particolare, la distribuzione delle aziende all'interno dei rispettivi settori è la seguente:

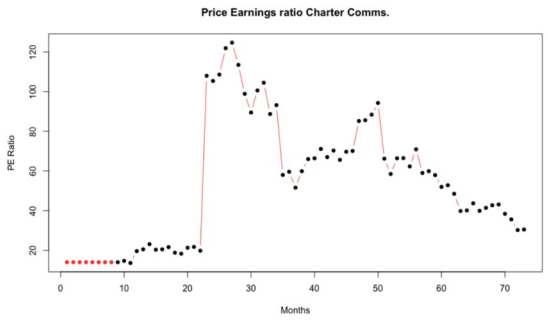
SETTORE	N° AZIENDE
COMMUNICATION SERVICES	15
CONSUMER DISCRETIONARY	46
CONSUMER STAPLES	26
ENERGY	2
FINANCIAL	21
HEALTH CARE	46
INDUSTRIAL	57
MATERIALS	19
REAL ESTATE	27
TECHNOLOGY	43
UTILITIES	22
TOT	324

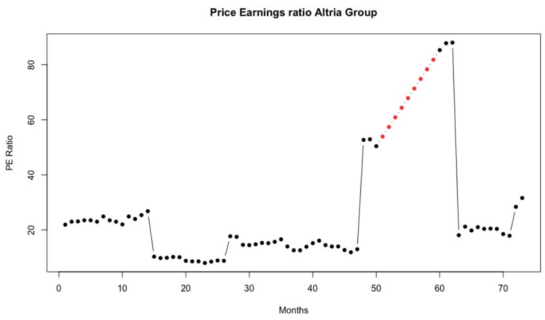


Si nota come la maggior parte delle aziende appartenga al settore *industrial*, seguito dai settori *technology, health care* e *consumer discretionary,* mentre il settore meno rappresentativo è quello energetico.

1 Nel seguito abbreviato in "S&P 500".

Non tutte le aziende considerate presentavano dati sufficienti allo svolgimento delle analisi: molte di esse mostravano molteplici dati mancanti per diversi indicatori di bilancio. Si è, dunque, deciso di eliminare dal dataset tutte le aziende che presentavano più del 20% di dati mancanti per almeno un indice di bilancio. Ciò ha condotto ad una riduzione del numero di aziende a 324.
Tra le rimanenti aziende, alcune presentavano ancora dei *missing values*, seppur in numero ridotto. La risoluzione di questo problema è stata approcciata con due diversi metodi, a seconda che il dato mancante si trovasse all'estremità (primo o ultimo dato) o all'interno della serie storica. Nel caso in cui i dati mancanti fossero all'inizio della serie storica, è stato preso il primo dato non mancante, ed è stato inserito al posto dei *missing values* precedenti. Se, invece, la serie storica terminava con un valore mancante, si è preso l'ultimo dato presente, e lo si è inserito al posto dei successivi *missing values*.
Mentre, se i valori mancanti erano disposti all'interno della serie storica, si è deciso di sostituirli utilizzando un'interpolazione lineare tra il valore precedente e il successivo alla sequenza di *missing values*.
Nei successivi grafici vengono rappresentate le modalità con cui è stata svolta la sostituzione dei *missing values*. In rosso sono rappresentati i dati precedentemente mancanti e successivamente sostituiti come descritto. Nel primo caso si avevano dati mancanti all'inizio, mentre nel secondo caso essi si trovavano all'interno della serie storica.





Una volta che il dataset è stato pulito da tutti gli eventuali valori mancanti, si è proceduto a rendere gli indicatori coerenti tra loro, questo perché in base a come un indicatore è costruito, un valore alto (o basso) può essere associato ad un aumento o ad una diminuzione del valore dell'azienda. Per soprassedere a questo problema, si è deciso di modificare il segno di alcuni indicatori, in modo tale che per tutti gli indici valga la regola secondo cui, tanto più è alto un indicatore, tanto più l'azienda è migliore. In questo modo, si è deciso di cambiare il segno di tutti gli indicatori di bilancio per i quali un valore alto è associato ad un'azienda peggiore.[2]
Nel seguito, per permettere di attuare un confronto tra indicatori di bilancio diversi, si è proceduto ad una standardizzazione degli stessi. La standardizzazione è stata svolta in questo modo: per ogni istante temporale e per ogni indicatore di bilancio, si è deciso di sottrarre ad ognuno di essi la sua media e dividerlo per la sua deviazione standard, in modo tale da disporre di dati sulla stessa scala e dunque confrontabili tra loro.

2 Nel seguito verranno specificati gli indicatori di bilancio per i quali è stato modificato il segno in modo da renderli coerenti tra loro.

1.2 Descrizione degli indici di bilancio

Una parte fondamentale di questo lavoro si basa sulla scelta e l'interpretazione degli indici di bilancio da utilizzare. In particolare, si è deciso di focalizzarsi su indicatori che andassero a valutare il *pricing* di un'azione, la sua redditività, la salute dell'azienda, le possibilità di crescita e la distribuzione di dividendi.
Le finalità di utilizzo degli indici di bilancio sono molteplici: essi permettono di "evidenziare gli aspetti più rilevanti della gestione e palesare collegamenti tra variabili economiche e finanziarie"[3], nonché di "facilitare confronti nel tempo […] e nello spazio"[4]. Gli indici di bilancio, quindi, forniscono una visione pragmatica dell'azienda, e verranno utilizzati con lo scopo di raggruppare società simili tra loro.
Di seguito si andrà a descrivere dettagliatamente tutti gli indicatori di bilancio utilizzati nell'analisi

- *1.2.1 Price Earnings Ratio:*

$$PE\ Ratio = \frac{Prezzo\ per\ azione}{Utile\ per\ azione}$$

Il *Price Earnings Ratio*[5] è il rapporto tra il prezzo per azione e l'utile per azione di un'azienda.
Lo si può interpretare come il numero di volte che il prezzo di un titolo incorpora gli utili.
Viene utilizzato per valutare il prezzo di un'azione in rapporto agli utili che l'azienda sottostate genera, nonché per misurare il valore di una società.
Gli investitori utilizzano il *PE ratio* per capire se un'azione è sopravvalutata o sottovalutata. Spesso esso viene utilizzato per confrontare aziende tra loro e per capire quale è più conveniente da comprare.
Intuitivamente, se a parità di prezzo per azione il PE è alto, significa che l'azienda sottostante genera un utile basso. Ciò significa che un'azione può essere vista come sottovalutata nel momento in cui il suo PE è basso.
Per rendere il *PE ratio* coerente con gli altri indicatori, si è deciso di modificarne il segno, così facendo, ad un valore alto dell'indice corrisponde un'azienda sottovalutata, ovvero un'azione che a parità di utile generato viene scambiata ad un prezzo più basso.

3 U. Sòstero, P. Ferrarese, Analisi di Bilancio, strutture formali, indicatori e rendiconto finanziario, Milano, Giuffrè Editore, 2000, p.60.
4 U. Sòstero, P. Ferrarese, Analisi di Bilancio, strutture formali, indicatori e rendiconto finanziario, Milano, Giuffrè Editore, 2000, p.61.
5 Nel seguito abbreviato in "PE" o "*PE ratio*".

- *1.2.2 Forward Price Earnings Ratio:*

$$Forward\ PE = \frac{Prezzo\ per\ azione}{Utile\ per\ azione\ atteso}$$

Il *Forward Price Earnings Ratio*[6] è una versione del *Price Earnings Ratio* nella quale al posto degli utili per azione correnti, vengono utilizzati gli utili per azione previsti per l'anno successivo. Perciò, rispetto al *PE ratio*, il *Forward PE* è una misura previsiva e non da informazioni certe rispetto al futuro dell'azienda.
Il *Forward PE* permette di stimare il valore futuro di una società sulla base delle previsioni sugli utili. Molti investitori utilizzano il *Forward PE* quando si tratta di valutare un'azienda, perché essi preferiscono fondare i loro investimenti su un'ottica di prospettiva futura della società, piuttosto che su un approccio retrospettivo.
Allo stesso modo del *Price Earnings ratio,* anche per il *Forward PE* si associa ad un basso valore dell'indicatore, un concetto di sottovalutazione dell'azienda da parte del mercato.
Anche in questo caso, perciò, si è deciso di modificare il segno dell'indice in modo tale da renderlo coerente con l'interpretazione degli altri indicatori, ovvero per fare in modo che un alto valore del *Forward PE* indichi un'azienda più sottovalutata.

- *1.2.3 Price to Book Value:*

$$PB\ Ratio = \frac{Prezzo\ di\ mercato}{Capitale\ proprio}$$

Il *Price to Book Value*[7] è il rapporto tra la quotazione di mercato e il capitale proprio per azione della società, anche detto valore contabile.
Il *PB ratio* viene utilizzato dagli investitori per capire se un'azione è prezzata coerentemente con il capitale proprio di cui la società sottostante dispone. In generale, le azioni che presentano un *PB ratio* basso sono considerate come sottovalutate dal mercato; al contrario, un elevato *Price to Book Valu*e indica che, fissato un certo valore contabile, l'azione viene scambiata a prezzi alti, e quindi meno vantaggiosi per un investitore.
Anche in questo caso, quindi, vengono considerate migliori quelle società che presentano un valore basso dell'indice. Per questo motivo si è di nuovo deciso di modificare il segno dell'indicatore in modo da associare un titolo migliore ad un valore dell'indice più alto.

6 Nel seguito abbreviato in "*Forward PE*" o "*Forward PE ratio*".
7 Nel seguito abbreviato in "PB" o "PB ratio".

- *1.2.4 Enterprise Multiple:*

$$Enterprise\ Multiple = \frac{Valore\ d'impresa}{EBITDA}$$

L'*Enterprise Multiple* è il rapporto tra l'*Enterprise Value* [8] e l'EBITDA[9] di una azienda.
L'*Enterprise Value* corrisponde al valore d'impresa, ovvero il valore di una società, mentre l'EBITDA rappresenta una misura di margine operativo lordo.
Questo indicatore viene utilizzato per determinare il valore economico di un'azienda e per capire se essa è sottovalutata o sopravvalutata.
L'*Enterprise Multiple* può variare molto in base al settore economico nel quale l'impresa opera. In particolare, esso viene confrontato con quello di aziende simili per paragonarle tra loro.
Generalmente, tanto più basso è l'*Enterprise Multiple*, tanto più un investitore considererà attraente l'azienda sottostante. Per questo motivo, quindi, si è deciso di modificare il segno di questo indicatore, in modo da rendere la sua interpretazione coerente con la decisione di associare un'azione migliore ad un quoziente più alto.

- *1.2.5 Return on Equity:*

$$ROE = \left(\frac{Reddito\ netto}{Capitale\ proprio} \right) * 100$$

Il *Return on Equity*[10] è tra i più importanti indicatori della redditività aziendale, ovvero quelli che si occupano di "valutare la capacità dell'impresa di produrre risultati economici soddisfacenti"[11].
Il ROE si calcola come il rapporto, espresso in percentuale, tra reddito netto e capitale proprio; per questo motivo viene anche chiamato indice di redditività del capitale proprio.
Questo indicatore esprime il rendimento complessivo dell'impresa dal punto di vista del portatore del capitale proprio, e riesce a fornire una sintesi dell'economicità della gestione aziendale, permettendo di valutare come il *management* sia riuscito a gestire il capitale proprio per aumentare gli utili.
Alla luce di ciò è evidente come un alto valore del ROE rispecchi una buona capacità da parte dell'impresa di far fruttare il capitale proprio, e di conseguenza una società più appetibile agli occhi di un investitore.

8 Nel seguito abbreviato in "EV".
9 EBITDA è l'acronimo per Earnings Before Interests, Taxes, Depreciations and Amortization.
10 Nel seguito abbreviato in "ROE".
11 U. Sòstero, P. Ferrarese, Analisi di Bilancio, strutture formali, indicatori e rendiconto finanziario, Milano, Giuffrè Editore, 2000, p.62.

- *1.2.6 Return on Assets:*

$$ROA = \left(\frac{Reddito\ operativo}{Attivo\ netto} \right) * 100$$

Il *Return on Assets*[12] è un indicatore molto usato nell'analisi della redditività aziendale. Il ROA indica quanto un'azienda è profittevole in relazione al totale degli asset a sua disposizione.
Questo indicatore viene utilizzato nell'analisi della gestione operativa di una società, la quale è volta a "valutare la capacità di produrre risultati economici soddisfacenti a partire da un determinato ammontare di risorse […]"[13].
Il ROA si calcola come il rapporto, espresso in percentuale, tra il reddito operativo e l'attivo netto; dove con attivo netto si intende il totale delle risorse finanziarie impiegate nella gestione dell'impresa.
Tale quoziente è, quindi, indicativo di quanto l'attivo netto dell'azienda venga utilizzato efficacemente. Ciò si traduce in un alto valore dell'indicatore quando la società è in grado di sfruttare le sue risorse finanziarie in modo profittevole, e in un basso valore dell'indicatore se la gestione operativa dell'impresa non è efficiente.

- *1.2.7 Earnings per Share:*

$$EPS = \frac{Utile\ netto}{N° azioni}$$

Gli *Earnings per Share*[14], o utili per azione, danno una misura dell'utile netto generato dalla società per ogni titolo azionario sul mercato.
Si calcola dividendo il profitto generato da un'impresa in un determinato periodo, per la media ponderata del numero di azioni in circolazione sul mercato in quello stesso arco temporale.
L'EPS viene utilizzato per determinare il valore attribuibile ad ogni singola azione di una società, ma anche come misura della redditività di un'azienda.
Generalmente gli investitori si aspettano che una società che presenta alti utili per azione abbia una buona capacità di generare ricchezza. Per questo motivo, a parità di altre condizioni, un'azienda con un EPS alto è preferibile ad una con utile per azione basso.

12 Nel seguito abbreviato in "ROA".
13 U. Sòstero, P. Ferrarese, Analisi di Bilancio, strutture formali, indicatori e rendiconto finanziario, Milano, Giuffrè Editore, 2000, p.65.
14 Nel seguito abbreviato in "EPS".

- *1.2.8 Dividend Yield:*

$$DY = \frac{Dividendo\ per\ azione}{Prezzo\ per\ azione} * 100$$

Il *Dividend Yield*[15] è dato dal rapporto, espresso in percentuale, tra il dividendo unitario pagato da una determinata azione e il prezzo dell'azione stessa.
Il DY è un indicatore finanziario che mostra quanto un'azienda paga in termini di dividendi, in relazione al prezzo al quale l'azione viene scambiata.
Più è elevato il *Dividend Yield*, migliore è il giudizio che viene espresso circa la capacità da parte della società di remunerare il capitale investito.
Spesso il DY varia in base alla dimensione dell'impresa e al settore all'interno del quale essa opera.
Dal momento che i dividendi partecipano attivamente all'aumento del rendimento di un investimento, a parità di condizioni, ci si aspetta che un investitore sia più propenso ad acquistare un titolo che paga dividendo maggiore.

- *1.2.9 Debt to Equity ratio:*

$$Debt\ Equity\ Ratio = \frac{Debito\ totale}{Capitale\ sociale}$$

Il *Debt to Equity ratio*[16] si calcola come il rapporto tra il debito e il capitale sociale di un'impresa in un certo istante.
Il DE è un indicatore che permette di valutare lo stato di salute di un'azienda, in quanto esso esprime il grado di indebitamento della stessa.
Esso fornisce una misura di quanto una società finanzi se stessa attraverso il debito piuttosto che attraverso il profitto. È evidente che un'azienda con un alto grado di indebitamento rispecchi una gestione non sana, mentre un'impresa che riesce a finanziarsi principalmente attraverso gli utili che genera, sia vista come più solida agli occhi di un investitore.
Anche in questo caso, quindi, si è deciso di modificare il segno dell'indicatore, in modo da associare ad un alto valore dell'indice un segnale positivo sull'azienda.

15 Nel seguito abbreviato in "DY".
16 Nel seguito abbreviato in "DE" o "Debt Equity".

- *1.2.10 Alcune statistiche descrittive:*

Di particolare interesse può essere la consultazione di alcune statistiche descrittive, che permettono di farsi un'idea di quali possono essere alcune quantità notevoli all'interno di ogni indicatore di bilancio.
Le statistiche descrittive scelte sono: media, mediana, deviazione standard, valore minimo e valore massimo.
Tutte queste quantità sono state calcolate per ognuno degli indicatori di bilancio, e in seguito riportati nella tabella sottostante.

	PE	FPE	PB	EM[17]	ROA	ROE	EPS	DY	DE
MEDIA	49.9	25.3	3.7	14.3	7.9	15.0	5.6	1.7	84.7
MEDIANA	25.6	19.7	3.9	12.7	6.8	15.9	3.5	1.5	76.9
DEV. STD.	451.8	63.9	69.6	6.8	7.6	619.2	12.6	1.5	1174.5
MIN	2.2	2.2	-2299.2	-3.3	-61.3	-24850	0.0	0.0	-77921.7
MAX	27700	4230	1120.7	85.0	70.6	6575.9	309.6	17.1	6036.7

1.3 Descrizione degli indicatori di rischio e rendimento

Oltre agli indicatori di bilancio, si è ritenuto fondamentale calcolare anche alcuni indicatori basati sui rendimenti delle singole azioni.
Per farlo si è partiti dalle serie storiche dei prezzi dei titoli giornalieri sull'arco temporale che va dal 31 dicembre 2014 fino al 31 dicembre 2021.
Innanzitutto, si è deciso di trasformare i prezzi in log-prezzi, e a partire da essi si sono calcolati i log-rendimenti[18] giornalieri, sottraendo al log-prezzo di ogni istante il log-prezzo dell'istante precedente.

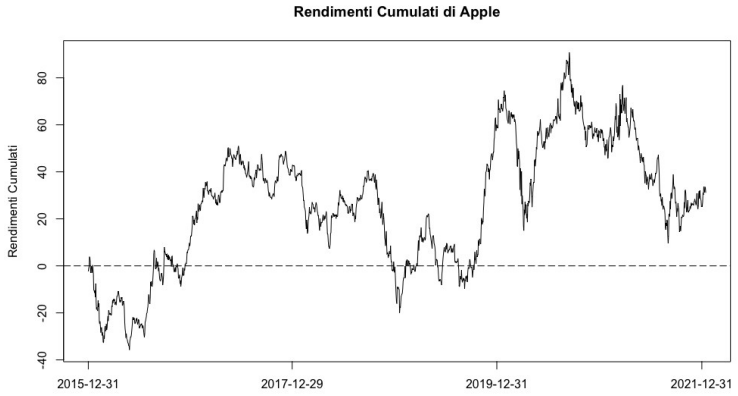
$$r_t = lp_t - lp_{t-1}$$

Una volta calcolata la serie storica dei rendimenti per ognuna delle 324 aziende appartenenti al dataset, si è potuto procedere con la costruzione di alcuni indicatori di rischio e rendimento come il Rendimento Cumulato, il Drawdown, e il Value at Risk.

17 Enterprise Multiple
18 Per semplicità, nel seguito ci si riferirà ai log rendimenti chiamandoli solo "rendimenti".

- *1.3.1 Il Rendimento Cumulato:*

Per calcolare il Rendimento Cumulato si è deciso di utilizzare un approccio a finestra mobile, con ampiezza della finestra pari ad un anno.
Questo indicatore è stato costruito per ognuna delle 324 aziende considerate con la seguente modalità: a partire dal 31 dicembre 2015 fino al 31 dicembre 2021, per ogni istante si è costruito un indicatore che andasse a sommare i rendimenti logaritmici dei 252 giorni precedenti[19].
Il Rendimento Cumulato è molto utile per farsi un'idea del guadagno in percentuale che un investimento su un particolare titolo potrebbe generare nell'arco di un anno.



Come esempio si può prendere il rendimento cumulato del titolo Apple.
Ad ogni istante esso mostra il rendimento cumulato del titolo nell'anno precedente. Quando il grafico va sotto lo zero significa che nell'ultimo anno i rendimenti negativi hanno superato i rendimenti positivi, mentre quando il grafico sta sopra lo zero, significa che negli ultimi 12 mesi il titolo ha aumentato la sua quotazione. Questo grafico permette di capire quali sono stati i periodi migliori e peggiori per un'azienda.

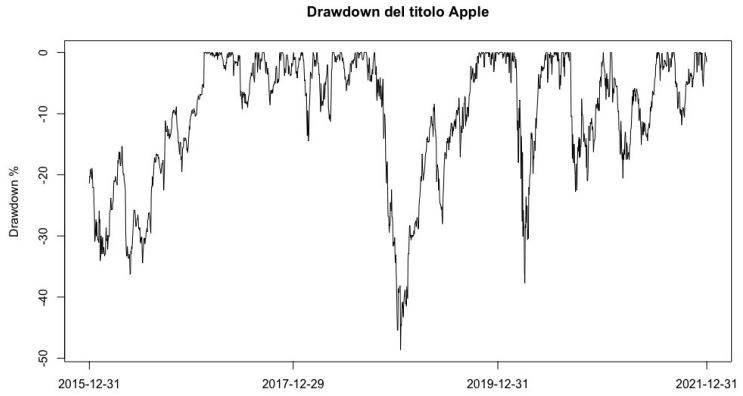
- *1.3.2 Il Drawdown:*

Il *Drawdown* esprime, in percentuale, la massima perdita di valore di un titolo dal picco di prezzo precedente.
Il *Drawdown* (D_t) di un titolo al tempo t si calcola ricorsivamente come il valore minimo tra zero e la somma tra il *Drawdown* (D_{t-1}) al tempo t-1 e il rendimento logaritmico dell'azione al tempo t (r_t).

19 Si è scelta una finestra di 252 giorni perché è il numero medio di giorni di apertura del mercato in un anno.

$$D_t = min\ (\ 0\ ,\ D_{t-1} + r_t\ )$$

Così facendo il *Drawdown* ad ogni istante può valere alternativamente 0, oppure un valore minore di zero, che corrisponde alla perdita massima che avrei potuto riscontrare se avessi investito nell'istante relativo all'ultimo massimo di prezzo. Conoscere il *Drawdown* è utile perché permette di disporre della correzione di prezzo massima che un'azione ha subito in un intervallo di tempo, conferendo un utile indicatore di rischio.

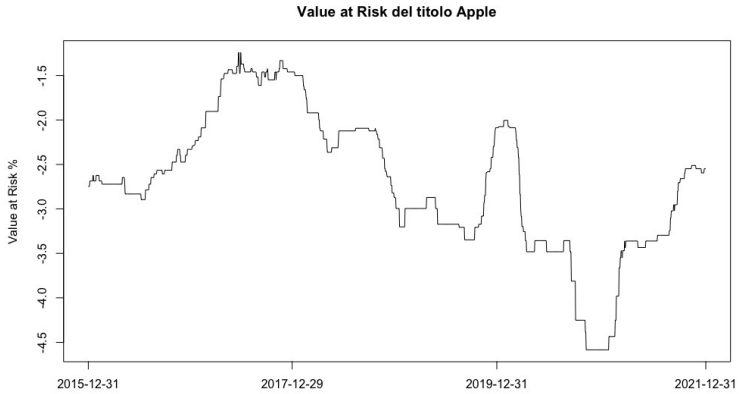


Come esempio si può utilizzare di nuovo il titolo Apple. Dal grafico del suo *Drawdown* si nota come attorno al 2016 ci sta stato un periodo in cui l'azienda è arrivata a perdere oltre il 30% dal massimo assoluto precedente. Allo stesso modo anche tra il 2018 e il 2019 c'è stato un periodo dove i prezzi sono rimasti al di sotto del massimo assoluto precedente, dove l'azione è arrivata a perdere anche più del 40% prima di tornare e superare il picco di prezzo precedente.

- *1.3.3 Il Value at Risk:*

Il *Value at Risk* [20] è un indicatore molto importante nell'ambito della valutazione degli investimenti. Fu sviluppato negli anni Novanta da J.P. Morgan, con l'obiettivo di sintetizzare i rischi derivanti da tutti gli investimenti a cui la banca era esposta.
Il VaR fornisce un'indicazione riguardo la massima perdita potenziale a cui può andare incontro un investimento, con un certo orizzonte temporale e con una certa probabilità.
È possibile calcolare il *Value at Risk* di un singolo titolo, ma anche di un portafoglio, tenendo conto delle correlazioni dei titoli con il mercato, della loro volatilità, dell'orizzonte temporale e dell'intervallo di confidenza.
In questo specifico caso, per calcolare il VaR si è deciso di procedere nel seguente modo: si è evitato di assumere distribuzioni teoriche per i rendimenti dei titoli, e si è andati a calcolare il quantile empirico al 5% della distribuzione dei rendimenti, presi con una finestra mobile di 252 giorni.

20 Nel seguito abbreviato in "VaR".



Prendendo come esempio ancora una volta il titolo Apple, si nota come il VaR oscilli tra un valore attorno al -1.5% e un valore del -4.5% circa.
È evidente come in corrispondenza di periodi di mercato ribassista, il rischio di perdite, ad un livello di confidenza del 95%, per gli investitori che detenevano il titolo Apple, fosse molto maggiore rispetto a periodi di mercato prevalentemente rialzista.

- *1.3.4 Alcune statistiche descrittive:*

Anche in questo caso vengono riportate delle quantità notevoli, utili per farsi un'idea degli indicatori sopra costruiti. Ancora una volta, vengono rappresentate la media, la mediana, la deviazione standard, il valore minimo e il valore massimo assunti dagli indicatori di rischio e rendimento quali, rispettivamente, i rendimenti cumulati, il *drawdown* e il *Value at Risk.*
La tabella sottostante sintetizza tali quantità:

	RCUM	DD[21]	VaR
MEDIA	14.1	-13.5	-2.5
MEDIANA	14.5	-7.9	-2.3
DEV. STD.	23.6	16.8	1.0
MIN	-148.4[22]	-181.6[23]	-8.5

21 Drawdown
22 Il valore minimo dei rendimenti cumulati è inferiore al -100%. Ciò è possibile in quanto vengono utilizzati i rendimenti logaritmici nel calcolo di tale indicatore.
23 Il valore minimo del *drawdown* è inferiore al -100%. Ciò è possibile in quanto vengono utilizzati i rendimenti logaritmici nel calcolo di tale indicatore.

MAX	209.0	0.0	-0.4

2 Analisi di Clustering

2.1 Il clustering

Il *clustering* è un insieme di metodi di analisi multivariata dei dati, che ha l'obiettivo di classificare degli elementi in gruppi in modo che questi ultimi siano più omogenei possibile. Questi gruppi, detti *cluster*, sono degli insiemi di oggetti che presentano delle similarità tra elementi dello stesso gruppo, e delle dissimilarità tra elementi di gruppi diversi. Generalmente, questa similarità viene associata ad una misura di distanza in uno spazio n-dimensionale, dove n è il numero di variabili che vanno a definire le coordinate del singolo elemento.
L'analisi di *clustering* risulta molto utile in quanto può suggerire interessanti ipotesi sulle relazioni che intercorrono tra i dati[24].
Esistono molteplici algoritmi di *clustering*: tra questi, uno dei più famosi è il metodo delle k medie, o algoritmo *k-means*, molto utilizzato per la sua semplicità e adattabilità.

2.2 Descrizione del metodo delle k medie

Il metodo delle k medie è un famoso algoritmo iterativo di *clustering*, molto efficiente quando si conosce a priori il numero di cluster ottimale. Gli elementi da fornire in input a tale algoritmo sono il numero di *cluster* "k" e i dati da classificare, i quali corrispondono essenzialmente a punti su uno spazio n-dimensionale.
Questo metodo permette di raggruppare gli oggetti in k *cluster* in base alle loro coordinate.
L'algoritmo si compone di tre semplici passi: il primo passo consiste nel definire un numero pari a "k" di centroidi, uno per ogni *cluster*. Il secondo *step* consiste nell'associare ogni elemento del *dataset* al *cluster* avente il centroide ad esso più vicino[25]. Il terzo è il passaggio che vede l'aggiornamento dei centroidi; qui vengono ricalcolati k nuovi centroidi come il baricentro dei *cluster* appena formati. Iterativamente si procederà ad una nuova assegnazione degli elementi ai gruppi appena aggiornati, sempre in modo tale che la distanza euclidea tra l'oggetto e il centroide sia minima.
Fondamentale per il funzionamento dell'algoritmo di clustering è la specificazione della tipologia di distanza da utilizzare. Esistono diversi tipi di distanze in uno spazio n-dimensionale, in questa analisi viene scelta quella più comune: la distanza euclidea, la quale si calcola come segue.

24 R. Johnson, D. Wichern, Applied Multivariate Statistical Analysis, Pearson Education, 6th Edition, 2014, p.671.
25 Essendo che nel caso trattato il clustering avviene su un dataset avente molte variabili, il concetto di "vicinanza" viene fatto coincidere alla distanza Euclidea tra due punti su uno spazio multidimensionale.

$$d_{ij}=\sqrt{\sum_{k=1}^{q}\left(x_{ik}-x_{jk}\right)^2}$$

Dove x_{ik} e x_{jk} con k = 1,…,q variabili su cui si svolge il clustering, sono i valori delle osservazioni i-esima e j-esima tra le quali si vuole calcolare la distanza[26].

Ripetendo in modo iterativo il passo due e tre, si avrà che i raggruppamenti continueranno a cambiare fino ad un certo punto in cui i gruppi saranno stabili; questo sarà il punto di terminazione dell'algoritmo.

Questo algoritmo presenta un vantaggio dovuto alla semplicità applicativa, tuttavia non si possono non considerare anche i suoi punti deboli.

Innanzitutto, il metodo delle k medie non ha una soluzione ottima univoca; ciò significa che ripetendo più volte l'algoritmo non è detto che esso abbia lo stesso risultato di classificazione.

In secondo luogo, nel caso in cui il numero di *cluster* non sia noto a priori, si deve trovare un metodo che permetta di definirlo in modo efficiente. Nell'analisi successiva si utilizzerà lo *screeplot* per capire che numero associare a k. Per farlo si applica l'algoritmo *k-means* diverse volte utilizzando un numero di gruppi crescente. Ad ogni applicazione si va a considerare la percentuale di varianza extra-gruppo che viene spiegata, e poi si rappresentano tutte queste quantità in un grafico a linea. Ci si aspetta che questo grafico presenti un "gomito", ovvero un punto a seguito del quale la varianza spiegata aumenta notevolmente. In corrispondenza di quel punto si troverà il numero di gruppi ottimale da utilizzare per applicare il *clustering*.

2.3 Clustering preliminare su due istanti notevoli

Prima di iniziare la vera e propria analisi di *clustering* su tutto l'orizzonte temporale a disposizione, si è preferito svolgere un'analisi preliminare su due periodi particolari e molto diversi tra loro, ovvero i mesi di dicembre 2019 e aprile 2020. Si sono scelti questi due periodi perché, pur essendo molto vicini tra loro, presentano due scenari completamente opposti: il primo relativo ad un mercato piuttosto tranquillo, non ancora scosso dalla notizia dello scoppio della pandemia, e il secondo relativo ad un mercato pieno di incertezze e di difficoltà, con le aziende costrette a fermarsi e con le loro quotazioni calate drasticamente.

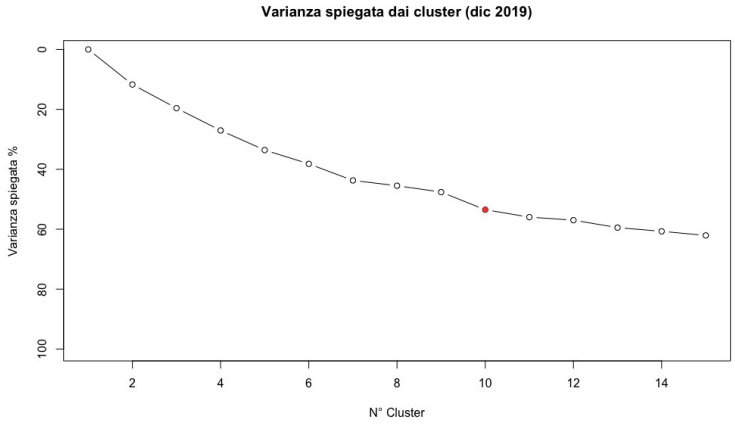
Questa analisi preliminare ha lo scopo di capire quali sono le differenze tra i *cluster* nei due periodi.

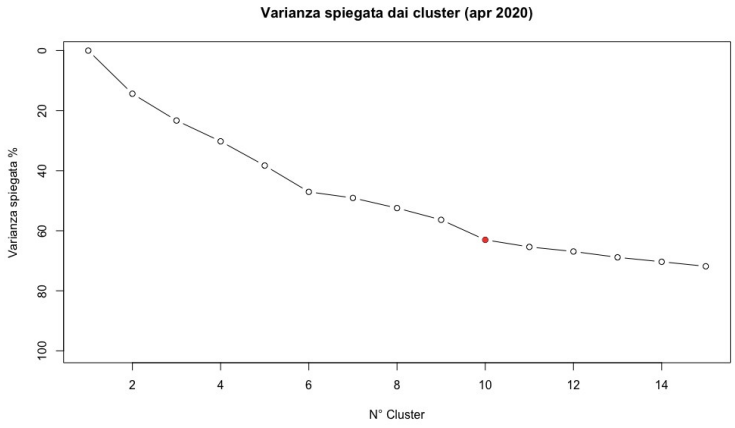
Come prima cosa si andrà a cercare di capire se il numero ottimale di gruppi cambia tra un periodo e l'altro. Dopodiché si andranno a valutare le varianze intra-gruppo, le medie intra-gruppo e i centroidi. Inoltre, si cercherà di capire come cambiano i *cluster* tra i due istanti.

A tale scopo, quindi, si sono costruiti due *data frame*: uno che contiene i valori degli indicatori di bilancio, rischio e rendimento a dicembre 2019, ed uno che contiene i valori degli stessi indicatori ad aprile 2020. Questi saranno i dati su cui si andrà ad applicare il *clustering*. Subito dopo, si sono costruiti dei grafici che permettono di scegliere il numero migliore di gruppi. L'approccio che è stato scelto è il medesimo che

26 B. Everitt, T. Hothorn, An introduction to applied multivariate analysis with R, Springer, 2011, p.15.

sta alla base del principio su cui si basa lo *screeplot*[27]: ovvero quello di guardare come varia la quota di varianza spiegata all'aumentare del numero di cluster. Rappresentando queste quantità, in un grafico a linea, ad un certo punto si dovrebbe notare una sorta di "gomito", ovvero il numero di *cluster* in seguito al quale la varianza spiegata non aumenta più in modo notevole.
In entrambi i casi considerati, i grafici evidenziano un gomito in corrispondenza del decimo cluster. Questo permette di mantenere il numero di gruppi costante e pari a 10 anche per l'analisi successiva.





Andando a svolgere il *clustering*, e confrontando il risultato con i settori ai quali appartengono le aziende, il risultato è il seguente:

Clustering per dicembre 2019

Cluster	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*1*	*Tot*

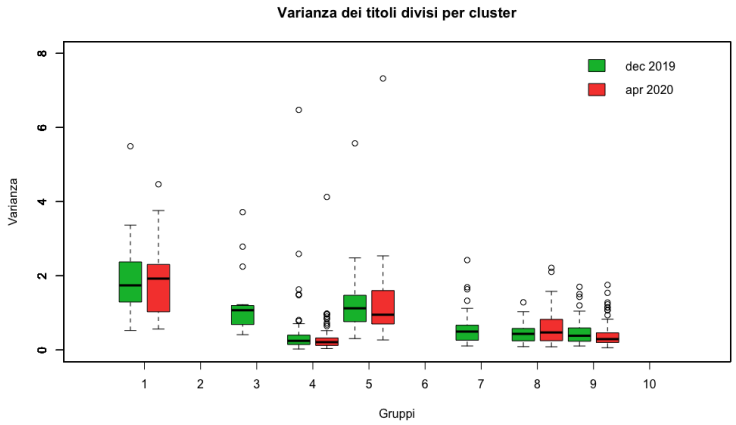
27 D. T. Christopoulos, Introducing Unit Invariant Knee (UIK) as an objective choice for elbow point in multivariate data analysis techniques, 2016.

										0	
Communication Services	2	0	0	4	1	0	0	0	2	6	15
Consumer Discretionary	4	0	0	14	3	2	8	9	4	2	46
Consumer Staples	2	0	0	9	1	1	1	0	12	0	26
Energy	2	0	0	0	0	0	0	0	0	0	2
Financial	2	0	0	9	1	2	4	2	1	0	21
Health Care	3	0	0	20	11	0	4	4	4	0	46
Industrial	2	0	0	27	1	2	8	8	9	0	57
Materials	1	0	0	8	0	0	3	5	2	0	19
Real Estate	0	3	1	0	0	0	1	20	2	0	27
Technology	2	1	16	1	4	0	0	1	2	0	43
Utilities	0	0	0	13	11	1	7	7	4	0	22
Tot	20	1	16	107	33	8	37	38	62	2	324

Clustering per aprile 2020

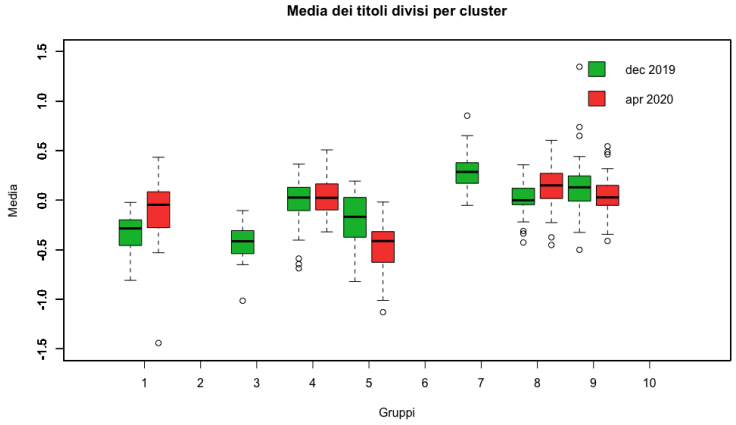
Cluster	***1***	***2***	***3***	***4***	***5***	***6***	***7***	***8***	***9***	***10***	***Tot***
Communication Services	0	4	3	1	5	0	0	0	0	2	15
Consumer Discretionary	1	18	1	3	3	0	2	0	2	16	46
Consumer Staples	0	5	2	0	16	0	2	0	0	1	26
Energy	0	0	2	0	0	0	0	0	0	0	2
Financial	0	9	1	2	4	0	1	0	0	4	21
Health Care	0	20	1	8	14	0	0	0	0	3	46
Industrial	0	26	0	0	13	0	1	0	0	17	57
Materials	0	7	0	0	4	0	0	0	0	8	19
Real Estate	0	3	7	9	6	0	0	1	0	1	27
Technology	0	1	0	6	5	6	8	0	17	0	43
Utilities	0	15	0	9	5	0	1	0	0	13	22
Tot	1	108	17	32	90	1	7	1	2	65	324

A questo punto, per cercare di capire come variano i *cluster* tra dicembre 2019 e aprile 2020, si è deciso di riordinare i gruppi di aprile 2020 in modo tale da renderli coerenti con quelli di dicembre 2019. Ciò significa, cercare di riordinare i *cluster* creati dall'algoritmo *k-means* in modo che al primo *cluster* di dicembre 2019 coincida il *cluster* di aprile 2020 che più gli è simile, e così via per tutti gli altri gruppi.
Per risolvere questo problema si è deciso di utilizzare una tabella a doppia entrata, dove sulle righe entrano i titoli divisi per cluster secondo la divisione di dicembre 2019, mentre sulle colonne entrano sempre i titoli divisi per *cluster*, ma con la suddivisione di aprile 2020.
L'idea è quella secondo cui, andando a massimizzare la diagonale di questa matrice, si riesca a capire quali sono i gruppi che più si somigliano nelle due suddivisioni. La risoluzione di questo problema porta ad avere che, ad ogni *cluster* del primo periodo, si possa associare il *cluster* del secondo periodo ad esso più simile, permettendo di fare confronti tra essi.
Una volta riordinati i *cluster*, si analizza la varianza intra-gruppo, andando a focalizzarsi sulle differenze tra i due istanti temporali e tra i diversi gruppi in uno stesso momento.
Nei grafici vengono tenuti in considerazione solo i *cluster* che presentano almeno 10 aziende, in quanto gli altri gruppi vengono considerati non significativi.
Per entrambi gli istanti, le aziende sono state separate tra loro in base al gruppo di appartenenza, dopodiché per ogni azienda è stata calcolata la varianza degli indicatori ad essa relativi per dicembre 2019 e per aprile 2020. Ci si è così trovati ad avere una misura di variabilità degli indicatori per ogni titolo, divisi per gruppo. Questi dati sono stati utilizzati per rappresentare i *boxplot* della varianza dei titoli divisi per *cluster*.

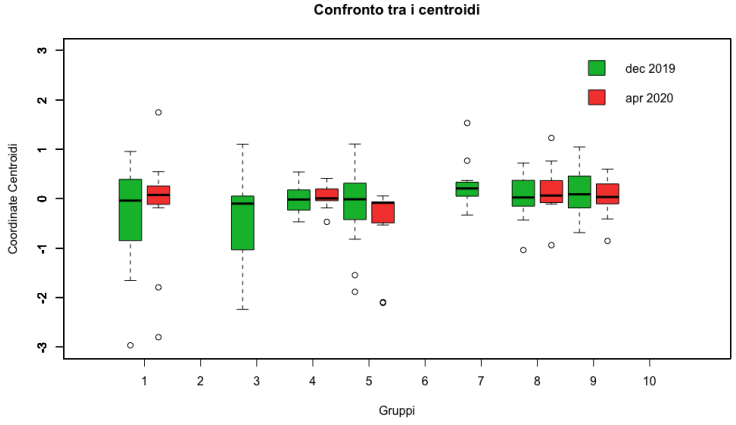


Come si può notare, le mediane delle varianze dei gruppi per entrambi gli istanti sono piuttosto eterogenee. Mentre se si va ad analizzare l'estensione dei *boxplot*, quindi la variabilità delle varianze dei titoli divisi per *cluster*, si nota che esse sono piuttosto simili quando si vanno a considerare i gruppi appaiati.
Si nota che i gruppi 4 e 9 in entrambi i periodi sono quelli con variabilità più ridotta e più simile tra un istante e l'altro. Mentre i *cluster* 1 e 5 presentano molta variabilità intra-gruppo per quanto riguarda le varianze degli indicatori dei titoli, sia a dicembre 2019 che ad aprile 2020.

Successivamente si è passati ad analizzare le medie dei *cluster*, sempre attraverso l'utilizzo dei boxplot.
La procedura è stata la stessa, ma al posto del calcolo della varianza degli indicatori per ogni titolo nei due istanti, qui è stata calcolata la media.



In questo caso è interessante notare come per alcuni gruppi la distribuzione delle medie degli indicatori sia simile tra un istante e l'altro, questo succede ad esempio per il *cluster* 4. Per altri gruppi invece, la differenza tra le medie è più evidente. Nel primo *cluster*, si ha un evidente aumento in media dei valori degli indicatori, si potrebbe pensare quindi che i titoli del primo gruppo abbiano reagito bene allo scoppio della pandemia. Per altri gruppi come il quinto si ha avuto una diminuzione degli indicatori in media. È possibile che questa volta, i relativi titoli abbiano reagito male agli effetti che il Covid-19 ha avuto sui mercati.
Successivamente, si è anche svolta un'analisi sui centroidi dei *cluster*. I centroidi sono dati dalle coordinate dei baricentri dei gruppi al termine dell'algoritmo *k-means*. In questo caso, avendo 12 variabili su cui viene svolto il *clustering*, ogni centroide è un punto in uno spazio a 12 dimensioni, dove ogni dimensione corrisponde ad uno degli indici di bilancio, rischio e rendimento.
Quindi ogni centroide è una collezione di 12 coordinate, le quali sono state rappresentate tramite *boxplot*, uno per ogni *cluster* nei due istanti considerati.



Questa analisi grafica permette di notare che le mediane delle coordinate di tutti i centroidi si aggirano intorno allo zero, indipendentemente dal fatto che i *cluster* appartengano a dicembre 2019 o ad aprile 2020. Non si evidenzia un *cluster* nettamente migliore degli altri sotto questo punto di vista, ma si può notare come, in genere, i centroidi dei *cluster* ad aprile 2020 abbiano una variabilità di coordinate più ridotta rispetto a quelli di dicembre 2019.
Per farsi un'idea di quale potrebbe essere il *cluster* migliore si può calcolare la mediana dei centroidi, e andare a scegliere il gruppo con la mediana più alta. Si preferisce la mediana alla media per evitare che valori estremi di alcune coordinate inducano a risultati fuorvianti.

CLUSTER	*Dicembre 2019*	*Aprile 2020*
1	-0.0372	**0.0769**
2	-0.4761	-0.0395
3	-0.0980	0.1041
4	-0.0159	0.0071
5	-0.0113	-0.0812
6	-0.0455	-0.0216
7	**0.2083**	0.1111
8	0.0262	0.0640
9	0.0902	0.0354
10	0.0121	0.0819

In base alla tabella sopra riportata, sempre escludendo i cluster con meno di 10 titoli, si può notare come, per il primo periodo il gruppo migliore sia il numero 7, mentre per il secondo periodo sia il numero 1.
In conclusione, svolgere questa analisi preliminare ha permesso di decidere che anche in seguito il numero di *cluster* che verranno utilizzati sarà 10. Inoltre, ha messo in luce alcune differenze e alcune similitudini tra il *clustering* svolto prima e dopo lo scoppio della pandemia ad inizio 2020.
Nel seguito si andrà a svolgere un'analisi simile di anno in anno, in modo da capire quanto e come cambiano i gruppi nel giro di un arco temporale di 12 mesi, e con l'obiettivo di costruire dei portafogli di titoli azionari sui *cluster* migliori.

2.4 Analisi di stabilità dei gruppi e stock picking

Questa sezione descrive uno dei punti focali del lavoro: si andrà a svolgere un'analisi di *clustering* alla fine di ogni anno, con l'obiettivo di farsi un'idea dei cambiamenti che

possono avvenire all'interno di un gruppo tra un anno e il successivo, ma anche con l'obiettivo di selezionare il *cluster* di aziende migliori ogni 12 mesi, e costruire con esse un portafoglio di titoli azionari per ogni anno.
Come prima cosa, quindi, si è andati a costruire 7 diversi *data frame*, uno alla fine di ogni anno tra il 2015 e il 2021, all'interno dei quali entravano tutte le azioni con i rispettivi indici di bilancio, rischio e rendimento standardizzati calcolati nel corrispondente istante temporale. Questi set di dati sono stati la base di partenza per applicare l'algoritmo di *clustering*. Ancora una volta il metodo utilizzato è stato il *k-means*, con k preso uguale a 10 in tutti gli istanti considerati.
Come già detto, uno dei punti deboli del metodo delle k medie è il fatto che l'ordine dei *cluster* varia tra un'applicazione dell'algoritmo e l'altra. Ciò significa che per analizzare la stabilità dei *cluster*, essi devono essere riordinati, in modo da risultare coerenti nel tempo, ovvero si è cercato di fare in modo che il k-esimo *cluster* al tempo t+1 fosse il più simile possibile al k-esimo *cluster* al tempo t.
Per risolvere questo problema si è utilizzata una tabella a doppia entrata, dove sulle righe entravano i gruppi al tempo t, mentre sulle colonne i gruppi al tempo t+1. L'idea è che andando a permutare le colonne, ovvero l'ordine dei *cluster* al tempo t+1, in modo da massimizzare la diagonale della matrice, si riesca a trovare un ordine per i gruppi al tempo t+1 in modo che essi siano i più coerenti possibile con i gruppi al tempo t.
Si consideri come esempio la tabella sotto riportata, essa presenta sulle righe i *cluster* del 2015 e sulle colonne i *cluster* del 2016 permutati in modo da massimizzare la diagonale.

CLUSTER	3	2	5	1	7	9	4	10	6	8
1	3	0	0	0	0	0	0	0	0	0
2	0	57	7	5	0	4	1	9	27	0
3	0	3	15	3	0	21	0	4	15	1
4	0	0	0	1	0	0	0	0	0	0
5	0	0	0	0	1	0	0	0	0	0
6	0	12	1	18	0	57	0	0	0	1
7	0	1	0	1	0	0	0	0	0	0
8	0	3	1	1	0	1	1	13	11	0
9	0	1	1	0	0	12	0	3	7	0
10	0	0	0	0	0	0	0	0	1	0

Così facendo si può dire che il terzo *cluster* del 2016 è, tra tutti, il più simile al primo *cluster* del 2015, che il secondo *cluster* del 2016 è il più simile al secondo *cluster* del 2015 e così via.

Se si applica questa procedura a tutti gli istanti considerati, si troverà che al k-esimo gruppo dell'anno 2015 corrisponde il k-esimo gruppo di tutti gli anni successivi.
Come esempio si possono considerare le due tabelle sottostanti, la prima riguarda i *cluster* nel 2015, mentre la seconda rappresenta i *cluster* nel 2016 riordinati in modo da risultare coerenti con quelli dell'anno precedente.

Cluster 2015

	1	2	3	4	5	6	7	8	9	10	TOT
COMMUNICATION SERVICES	0	4	1	1	0	5	0	3	0	1	15
CONSUMER DISCRETIONARY	2	20	4	0	1	7	1	5	6	0	46
CONSUMER STAPLES	0	10	1	0	0	14	0	0	1	0	26
ENERGY	0	0	1	0	0	0	0	0	1	0	2
FINANCIAL	0	7	4	0	0	8	1	0	1	0	21
HEALTH CARE	1	23	9	0	0	5	0	8	0	0	46
INDUSTRIAL	0	16	18	0	0	10	0	6	7	0	57
MATERIALS	0	6	8	0	0	1	0	2	2	0	19
REAL ESTATE	0	3	5	0	0	18	0	0	1	0	27
TECHNOLOGY	0	21	7	0	0	4	0	7	4	0	43
UTILITIES	0	0	4	0	0	17	0	0	1	0	22
TOT	3	110	62	1	1	89	2	31	24	1	324

Cluster 2016

	1	2	3	4	5	6	7	8	9	10	TOT
COMMUNICATION SERVICES	0	2	0	1	0	5	0	2	4	1	15
CONSUMER DISCRETIONARY	2	12	2	1	1	11	1	8	8	0	46

CONSUMER STAPLES	0	11	0	0	0	12	0	2	1	0	26
ENERGY	0	0	1	0	0	0	0	0	1	0	2
FINANCIAL	0	6	3	1	0	6	0	2	3	0	21
HEALTH CARE	1	19	5	1	0	6	0	9	5	0	46
INDUSTRIAL	0	18	4	0	0	16	0	2	16	1	57
MATERIALS	0	2	3	0	0	7	0	0	7	0	19
REAL ESTATE	0	0	1	23	0	2	0	1	0	0	27
TECHNOLOGY	0	7	6	2	0	8	1	3	16	0	43
UTILITIES	0	0	0	0	0	22	0	0	0	0	22
TOT	3	77	25	29	1	95	2	29	61	2	324

Avere dei *cluster* coerenti tra loro nel tempo sarà fondamentale per svolgere le analisi successive e per valutare la stabilità dei gruppi.
In questo senso, si è deciso di utilizzare un test del Chi-quadro sulle tabelle a doppia entrata, con l'obiettivo di farsi un'idea di quanto i gruppi cambino tra un anno e il successivo.
Il test del Chi-quadro su tabelle di contingenza permette di misurare la dipendenza tra le due variabili qualitative che entrano rispettivamente sulle righe e sulle colonne della tabella. In questo modo è possibile capire quanto i gruppi all'istante t+1 siano dipendenti dai gruppi al tempo t, e di conseguenza farsi un'idea di quanto i cluster rimangano stabili nel tempo.
Il test del Chi-quadro per tabelle di contingenza si costruisce come segue:

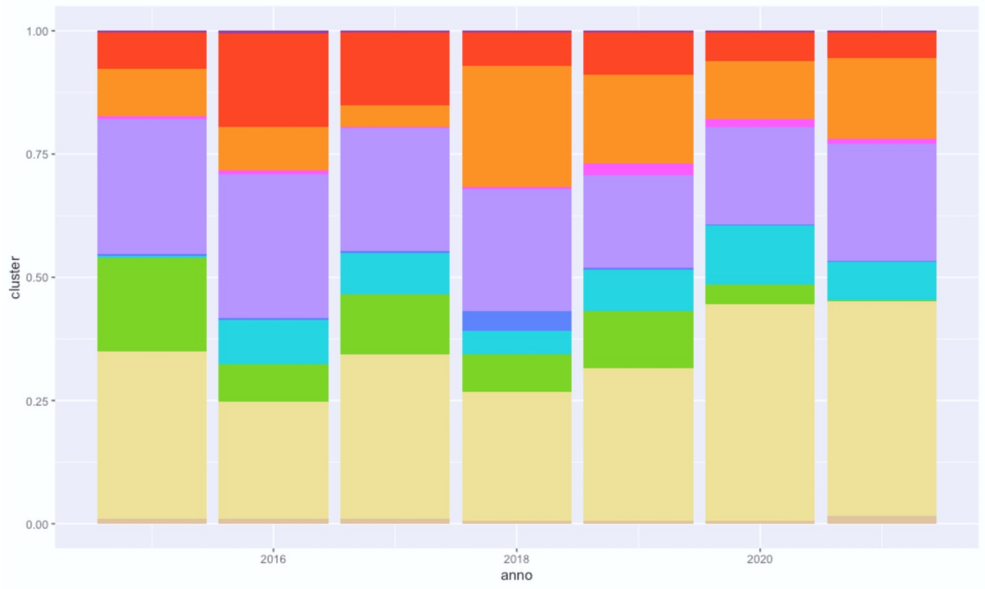
$$X^2 = \sum_{i=1}^{n} \frac{(O_i - A_i)^2}{A_i}$$

Dove O_i e A_i sono rispettivamente le frequenze osservate e le frequenze attese nella tabella di contingenza.
Andando a svolgere il test su tutte le tabelle a doppia entrata costruite come specificato in precedenza, i risultati sono i seguenti:

TABELLE	STAT TEST	P-VALUE
2015 - 2016	923.5985	<0.0001

2016 - 2017	865.951	<0.0001
2017 - 2018	579.8442	<0.0001
2018 – 2019	717.7258	<0.0001
2019 - 2020	343.6026	<0.0001
2020 - 2021	974.9628	<0.0001

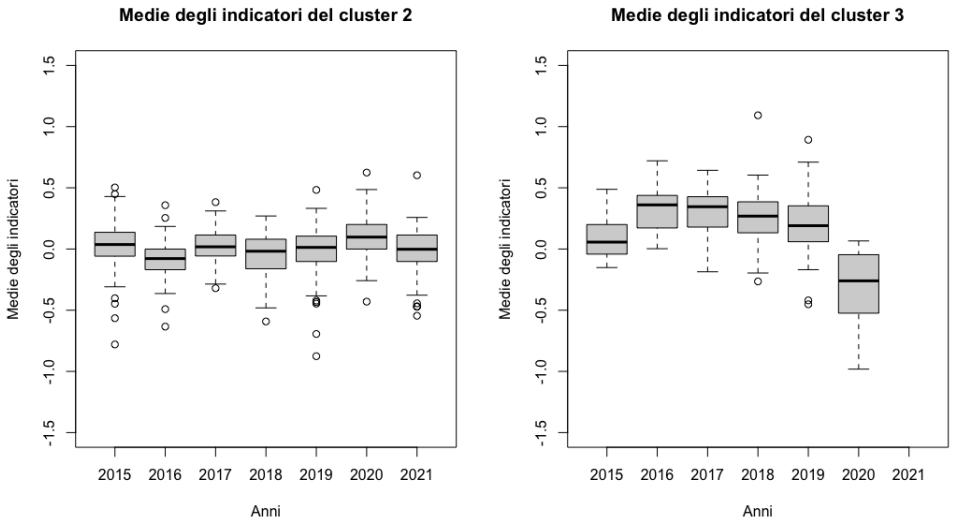
La tabella precedente riporta le statistiche test e i *p-value* dei test del Chi-quadro svolti sulle rispettive tabelle a doppia entrata. È ben visibile come non ci sia evidenza di indipendenza tra gruppi in due anni successivi, ma anzi, i dati supportano fortemente l'ipotesi di dipendenza, presentando statistiche test con valori molto alti e conseguentemente *p-value* molto bassi.
Si può, quindi, dire che i gruppi sono piuttosto stabili nel tempo.
Il seguente grafico ad area permette di visualizzare come cambiano le numerosità dei singoli gruppi anno per anno. Ogni colonna si riferisce ad un anno, e i diversi colori rappresentano ognuno dei 10 *cluster*.
È evidente come, salvo per alcuni casi, la numerosità di uno stesso gruppo non vede cambiamenti radicali nel tempo. Si può, quindi, affermare che i gruppi sono relativamente stabili anno per anno dal punto di vista della numerosità.

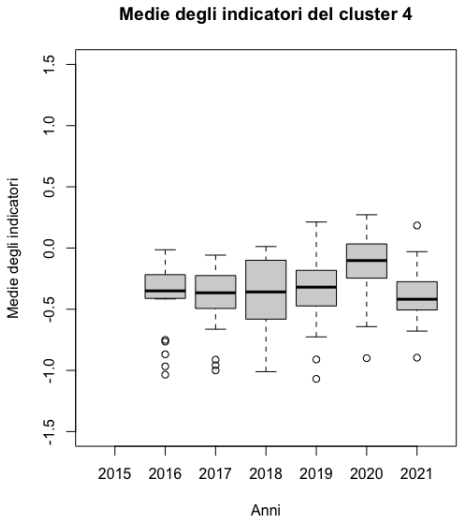


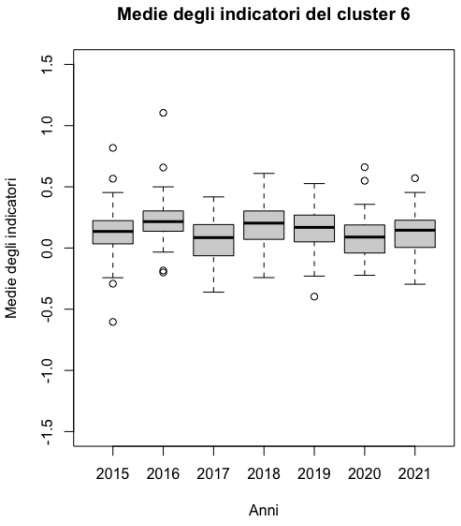
2.4.1 Analisi delle medie degli indicatori

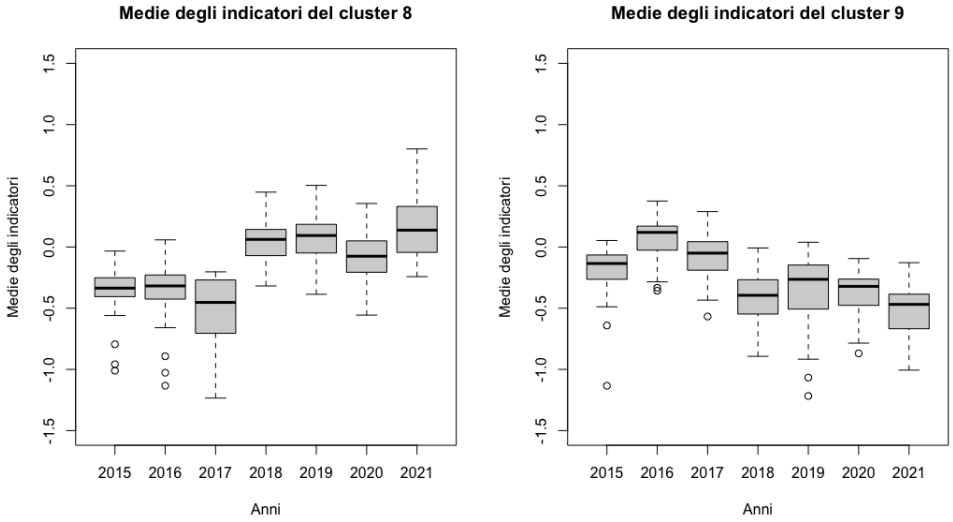
In seguito, si è potuto andare ad analizzare come variano i *cluster* di anno in anno dal punto di vista della media degli indicatori di bilancio, rischio e rendimento.
Per fare questo si sono calcolate le medie degli indicatori standardizzati per ogni titolo e in ogni istante considerato; queste quantità sono poi state ripartite nei corrispettivi *cluster*, in modo da poter fare confronti tra essi.

Successivamente, è possibile analizzare come variano le medie degli indicatori nei gruppi corrispondenti tra un anno e l'altro utilizzando un *boxplot*.
Nelle successive analisi sono stati rimossi i *cluster* non significativi, ovvero i gruppi che contenevano meno di 10 titoli. Tra il 2015 e il 2021, per tutti (o quasi) gli anni sono risultati non significativi i cluster 1,5,7,10. Nel 2015 risultava non significativo anche il cluster 4, mentre per il 2021 il cluster 3.







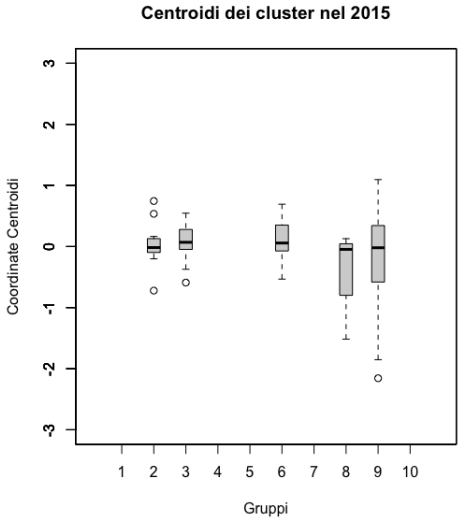


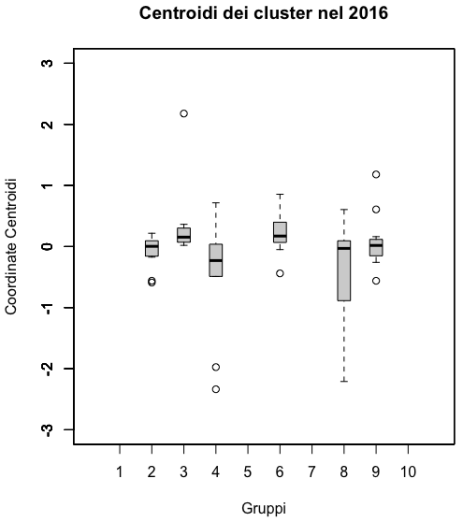
Analizzando *cluster* per *cluster*, si può notare come il gruppo 2 tenda ad avere una media degli indicatori piuttosto stabile nel tempo, e sempre attorno a zero, con una variabilità costante e contenuta.
Il terzo *cluster*, dopo un miglioramento tra l'anno 2015 e il 2016, vede un peggioramento nelle medie degli indicatori, in particolare tra la fine del 2019 e la fine del 2020, dove, tra le altre cose, aumenta anche la dispersione dei valori assunti dalle medie degli indicatori. Tuttavia, le mediane sono sempre positive, a parte per gli anni 2020 e 2021.
Il quarto gruppo non vede mai la mediana delle medie degli indicatori superare lo 0.
A parte per un miglioramento nell'anno 2020, non sembra che questo *cluster* abbia un'evoluzione positiva o negativa degli indicatori di bilancio, rischio e rendimento. Al contrario sembrano titoli con indicatori mediamente stagnanti e non particolarmente positivi.
Anche il sesto *cluster* sembra non presentare particolari aumenti o diminuzioni delle medie degli indicatori, tuttavia rispetto al quarto gruppo, qui le mediane sono sempre maggiori di zero, e la loro dispersione è più ridotta.
Il gruppo numero 8 vede un repentino miglioramento degli indicatori in media tra il 2017 e il 2018, dovuto non tanto ad un miglioramento delle aziende nel corso di un anno, ma piuttosto ad una rotazione dei titoli appartenenti a questo gruppo tra un anno e il successivo.
Il nono gruppo, invece, presenta una chiara tendenza peggiorativa per quanto riguarda le medie degli indicatori. Anche qui, a parte per il 2016, le mediane non superano mai lo zero, indicando aziende poco appetibili rispetto a quelle di altri cluster.

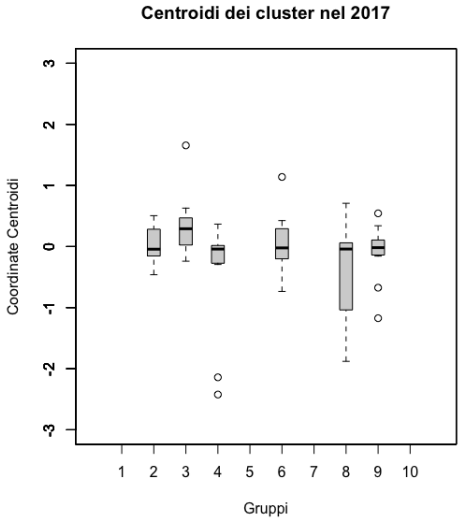
2.4.2 Analisi dei centroidi

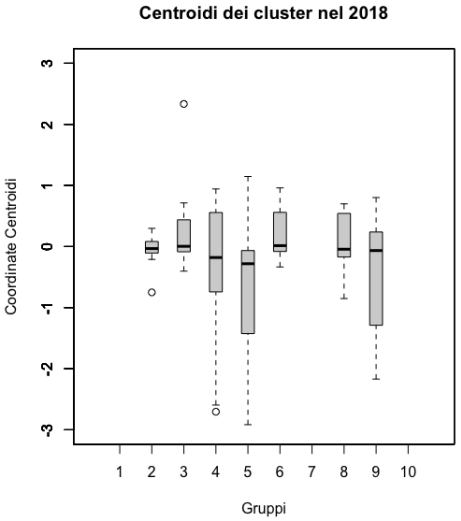
Oltre alle medie degli indicatori, potrebbe essere utile concentrarsi anche sui centroidi dei *cluster*. I centroidi danno un'idea di dove si collochi il baricentro di un gruppo, fornendo una misura sintetica che permette di confrontare i diversi *cluster* tra loro.
Come già detto, ogni centroide è una collezione di 12 coordinate, dove 12 è il numero di indicatori usati per svolgere l'analisi di *clustering*; è quindi possibile farsi un'idea di

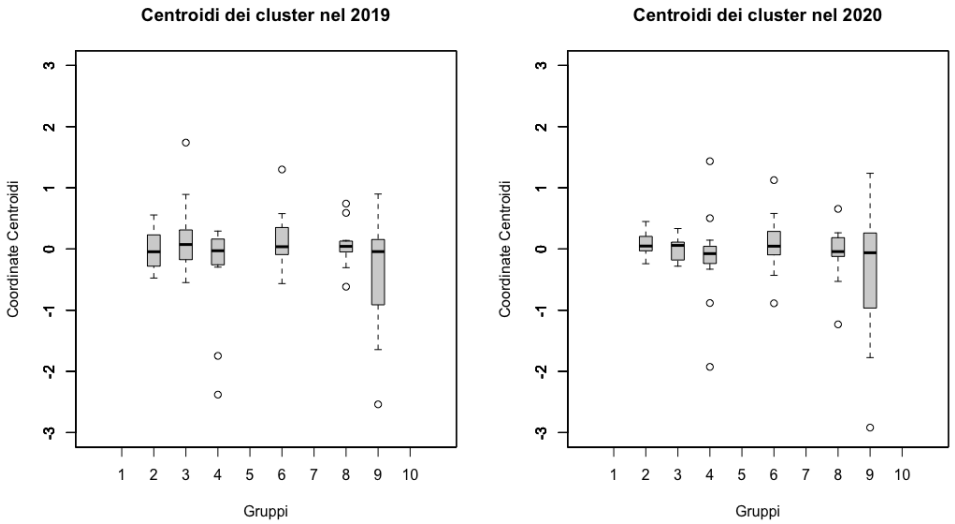
come variano i gruppi in uno stesso istante guardando i boxplot delle coordinate dei singoli centroidi per ogni anno.

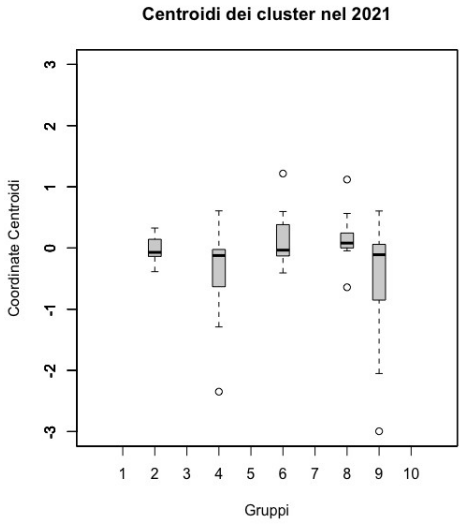






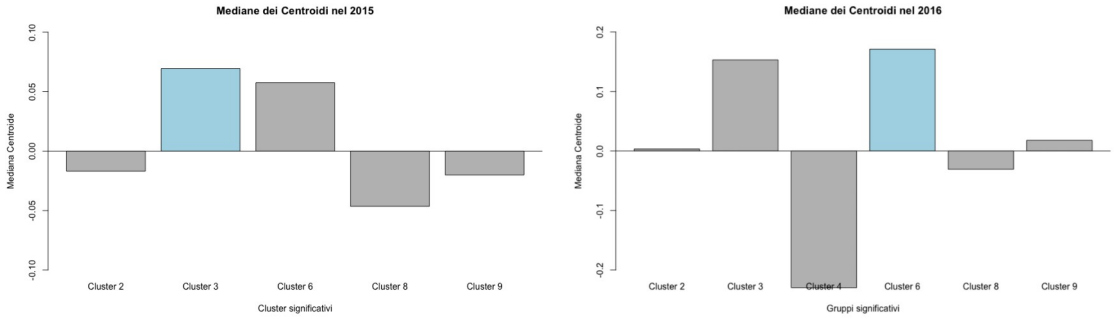


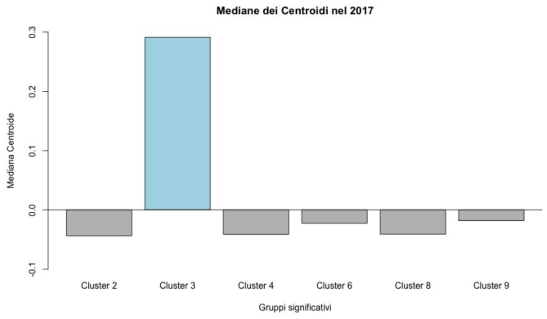


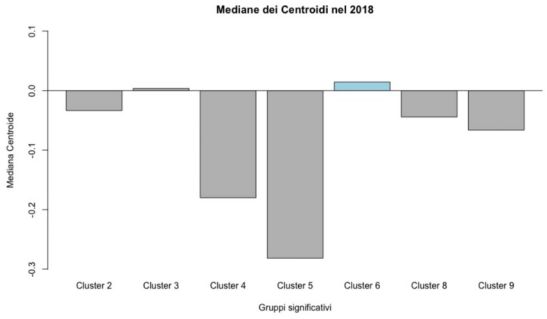


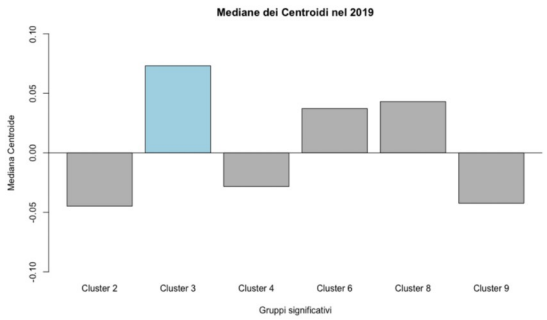
I *boxplot* permettono di visualizzare la posizione della mediana delle coordinate dei centroidi, ma anche di valutarne la variabilità.
La variabilità delle coordinate dei centroidi permette di capire quanto sono concentrati i titoli di un gruppo attorno al baricentro; infatti, un *boxplot* poco ampio indica che i titoli sono prevalentemente concentrati attorno al centroide, e quindi denota un gruppo di aziende più omogenee tra loro dal punto di vista degli indicatori utilizzati.
Si nota, ad esempio, come il centroide del gruppo 8 abbia coordinate molto eterogenee negli anni dal 2015 al 2018, mentre il centroide del gruppo 9 nell'anno 2015 e dal 2018 al 2021. Gli altri centroidi hanno generalmente coordinate meno variabli nel tempo, in particolare quello del gruppo 2.
Tuttavia, la cosa più interessante su cui focalizzarsi è la mediana delle coordinate dei centroidi. Essa sintetizza quello che potrebbe essere visto come un valore atteso robusto delle coordinate dei baricentri dei singoli *cluster*, permettendo un confronto a livello di gruppi su uno stesso istante temporale.

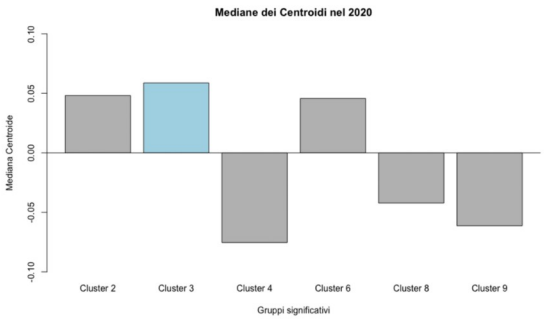
Secondo questo ragionamento, quindi, si è deciso di confrontare le mediane delle coordinate dei centroidi relativi ai *cluster* di uno stesso anno, con l'obiettivo di selezionare, per ogni anno, il *cluster* con il valore più alto.
Sotto questo punto di vista, e sotto i punti di vista secondo cui si è creato il *dataset*, un valore mediano alto delle coordinate di un centroide indica un gruppo di titoli migliore.
A questo punto si è potuta svolgere un'analisi grafica delle mediane delle coordinate dei centroidi anno per anno tramite l'utilizzo di un *barplot*. Si è deciso di omettere dal *barplot* i centroidi relativi ai *cluster* non significativi, per evitare di essere condotti ad una selezione inconsistente del gruppo migliore.

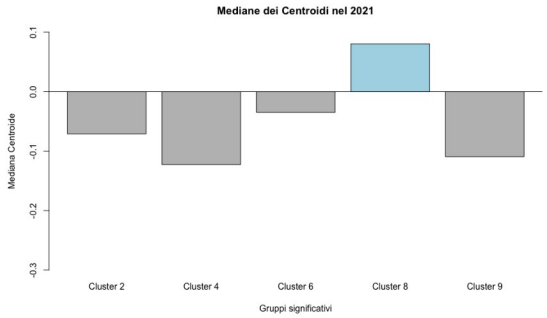












Secondo questo criterio di scelta, i *cluster* migliori per ogni anno sono quelli evidenziati in nei grafici.
È interessante notare come i *cluster* migliori siano sempre il terzo o il sesto, ad eccezione di fine 2021, dove il miglior gruppo è l'ottavo.
Tra il 2015 e il 2020, l'unico *cluster* ad avere sempre la mediana delle coordinate del centroide superiore a zero è proprio il terzo gruppo, evidenziando una buona solidità nelle aziende che lo compongono. Il *cluster* numero quattro, invece, è l'unico ad avere sempre una mediana delle coordinate del centroide inferiore a zero, indicando un gruppo di aziende non molto buone dal punto di vista degli indicatori considerati.
Una volta selezionato quello che dovrebbe essere il *cluster* migliore per ogni anno, l'obiettivo è valutare se l'investimento nei relativi titoli azionari sia conveniente o meno.

3 Applicazione dei risultati e conclusione

Nel precedente capitolo è stato implementato un metodo che permette di selezionare in modo semi-automatico un sottogruppo di azioni a partire da un paniere più ampio, con lo scopo di andare a cogliere le aziende migliori sotto il punto di vista degli indicatori utilizzati.
Ora l'obiettivo è quello di testare a posteriori se l'applicazione di questo metodo avrebbe portato ad un guadagno o ad una perdita, e di trarre alcune conclusioni sull'analisi svolta.

3.1 Analisi delle performance dei portafogli

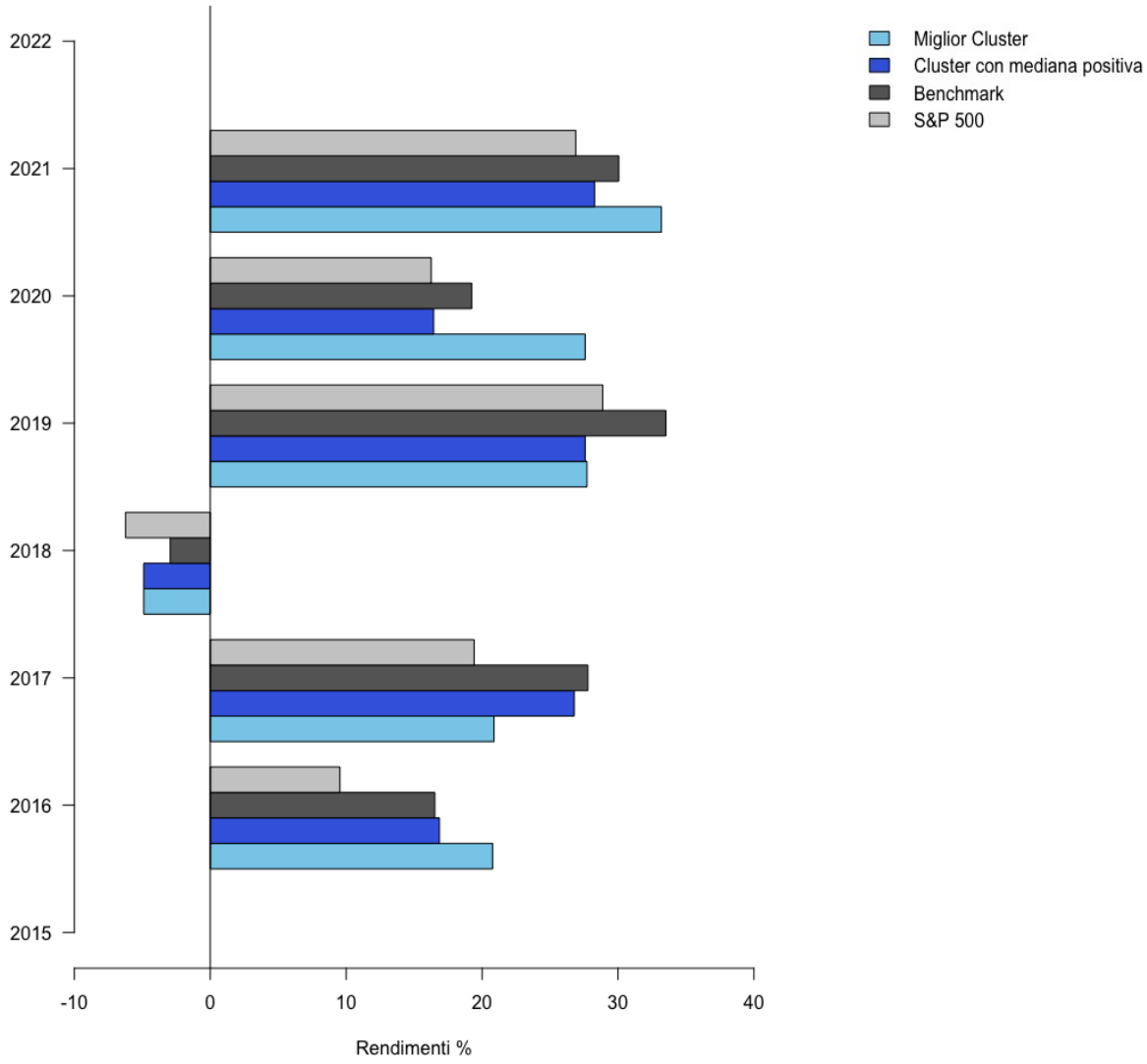
Una volta che l'analisi di *clustering* è stata svolta e che è stato possibile selezionare il miglior gruppo di aziende di anno in anno, si è voluto vedere se effettivamente questo metodo avrebbe portato ad un guadagno o meno.
Per fare ciò si è deciso di costruire, con ognuno dei *cluster* selezionati alla fine di ogni anno, un portafoglio *equally weighted*, e di confrontare le sue performance sull'anno successivo con le performance di altri tre portafogli: un cosiddetto "*Benchmark*" che consiste in un portafoglio equi-pesato composto di tutti i 324 titoli su cui è stata svolta l'analisi di *clustering*, un secondo portafoglio che replica l'S&P 500, e un terzo portafoglio costruito ancora una volta come portafoglio *equally weighted* nel quale entrano i titoli appartenenti ai *cluster* che alla fine di quello specifico anno presentavano una mediana delle coordinate dei gruppi maggiore di zero.
Con portafoglio *equally weighted* si intende un portafoglio di asset finanziari, dove ognuno di essi assume lo stesso peso, pari a 1/N, dove N è il numero di titoli che fanno parte del portafoglio.
Alla fine di ogni anno, quindi, è stato scelto il gruppo di aziende migliore, con le quali è stato costruito un portafoglio *equally weighted*. Su questo portafoglio è stato calcolato il

rendimento [28] che esso avrebbe ottenuto nell'anno successivo, e, subito dopo, confrontato con il rendimento che avrebbero ottenuto gli altri tre portafogli nello stesso arco temporale.
Il rendimento di un portafoglio, in generale, viene calcolato come

$$R_p = \omega' * R_i$$

Dove R_p è il rendimento del generico portafoglio, ω' è il vettore trasposto dei pesi con cui i singoli titoli entrano nel portafoglio, e R_i sono i rendimenti delle singole azioni.
Ad esempio, alla fine del 2015 è stato scelto il terzo *cluster*, i cui titoli sono stati utilizzati per costruire un portafoglio sul quale si è immaginato di aver investito per tutto il 2016. Si è, quindi, calcolato il rendimento di questo portafoglio tra l'1 gennaio 2016 e il 31 dicembre 2016, e lo si è confrontato con il rendimento del *Benchmark*, dell'S&P 500 e del portafoglio contenente tutti i *cluster* con mediana maggiore di zero. Svolgendo lo stesso procedimento per ogni anno, il risultato trovato è stato il seguente:

28 In questo caso è stato deciso di utilizzare i rendimenti, e non i log-rendimenti, in modo da rendere i risultati più "tangibili" e coerenti con i guadagni o perdite effettivi che si sarebbero realizzati.

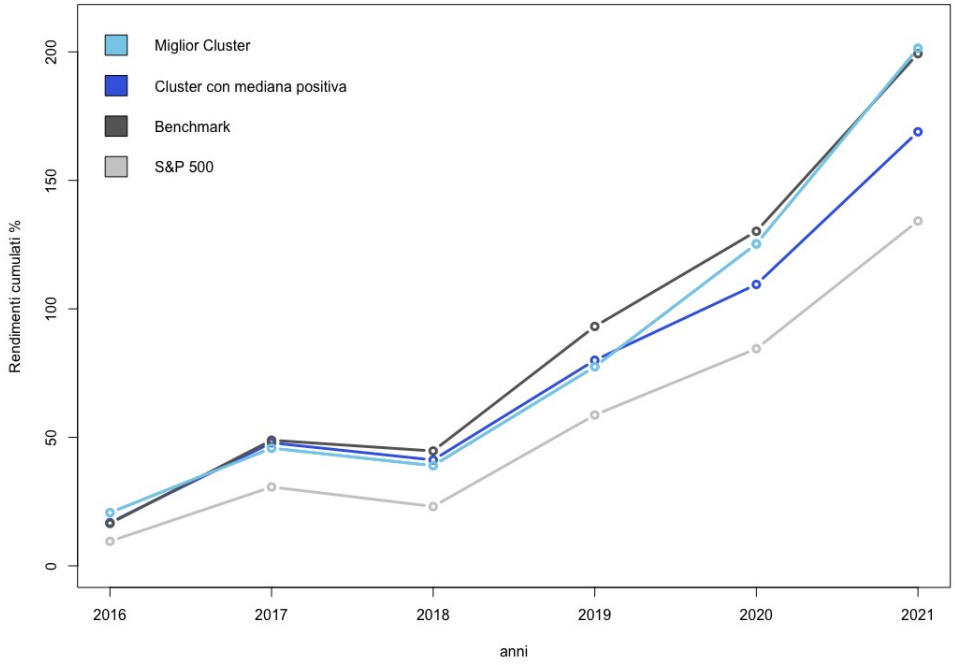


Confrontando i rendimenti del portafoglio costruito con i titoli appartenenti al *cluster* migliore e i rendimenti del portafoglio costruito con i titoli appartenenti a tutti i gruppi con mediana delle coordinate dei centroidi maggiore di zero, si nota come il secondo sovraperformi il primo solo in 1 anno su 6. Ciò evidenzia come il criterio di scelta del portafoglio sia efficiente, e porti alla selezione di un gruppo di aziende con performance piuttosto buone.
Andando a confrontare i rendimenti del portafoglio costruito sul *cluster* migliore con quelli dell'S&P 500 è evidente come, a parte per il 2019, essi siano sempre superiori. Questo risultato è, ancora una volta, molto positivo, e permette di affermare con relativa confidenza che il portafoglio selezionato tramite la precedente analisi di *clustering* sia migliore, dal punto di vista dei rendimenti, di un eventuale fondo che investe in modo passivo sull'S&P 500.
Se invece si confrontano i rendimenti del portafoglio relativo al miglior gruppo con i rendimenti del portafoglio *Benchmark*, si nota come 3 volte su 6 le aziende selezionate

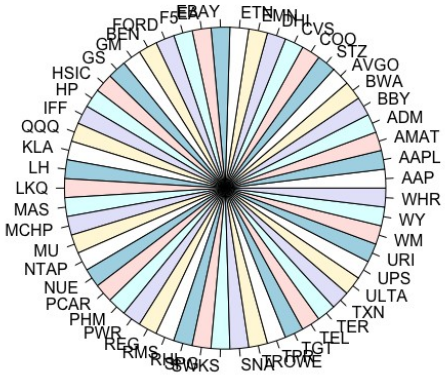
tramite l'analisi di *clustering* siano riuscite a battere l'insieme di tutte aziende considerate.
Questo risultato non permette di dire quale dei due approcci sia migliore, per cui si è deciso di considerare anche un'analisi dei rendimenti cumulati che questi portafogli avrebbero permesso di ottenere tra il 2016 e il 2021.
I rendimenti cumulati sono stati calcolati nel seguente modo:

$$Rcum_t = \{\prod_{i=0}^{t} (1 + R_{t-i}) - 1\} * 100$$

Dove R_{t-i} sono i rendimenti del portafoglio calcolati all'istante t–i.



Come è evidente, i rendimenti cumulati dell'S&P 500 sono inferiori rispetto a quelli di tutti gli altri portafogli considerati, fermandosi ad un profitto del 134.2% in 6 anni.
I portafogli costruiti con i titoli appartenenti ai *cluster* con mediana delle coordinate maggiore di zero, presentano un rendimento complessivo migliore dell'S&P 500, pari al 168.9%, tuttavia, essi non risultano essere allo stesso livello, dal punto di vista dei rendimenti cumulati, dei due portafogli seguenti.
Il portafoglio *Benchmark* avrebbe permesso un rendimento cumulato pari al 199.3%, un paio di punti percentuali più basso rispetto a quello dei portafogli selezionati, i quali avrebbero permesso un profitto pari al 201.4% in 6 anni: un risultato più che dignitoso in termini assoluti, ma piuttosto buono anche quando confrontato con gli altri portafogli considerati.
In tutto ciò rimane fuori l'analisi delle performance del portafoglio costruito alla fine del 2021, i cui rendimenti realizzati saranno noti solo alla fine del 2022. Il portafoglio in questione, tuttavia, è così composto:



Sarà interessante tornare ad analizzare il suo rendimento alla fine dell'anno per trarre qualche conclusione in più sull'efficienza del metodo di *screening* utilizzato.

3.2 Conclusioni

Al fine di trarre delle conclusioni, si è ripercorso il lavoro dal principio con l'obiettivo di vedere quali sono stati gli aspetti più degni di nota, cosa sarebbe stato possibile migliorare e cosa, invece, è risultato positivo.
Tutto sommato il percorso può considerarsi buono. Gli obiettivi, che erano quelli di riuscire a farsi un'idea di come variano i gruppi di titoli nel tempo, e di costruire dei portafogli di investimento con essi, sono stati raggiunti. In particolare, si è notato che in generale i gruppi rimangono piuttosto stabili nel tempo; ma soprattutto si è riusciti ad individuare, per ogni istante temporale considerato, un gruppo di titoli migliore degli altri, con il quale si è potuto costruire un portafoglio, che ha quasi sempre sovraperformato l'S&P 500 dal punto di vista del rendimento.
I risultati, quindi, sono stati soddisfacenti, ma potrebbero esserlo stati anche di più se si fossero scelti indicatori in modo più oculato, e in numero più ampio.
Inoltre, sarebbe stato possibile aumentare l'affidabilità dei risultati considerando un maggior numero di istanti temporali su cui svolgere il *clustering*.
Tuttavia, malgrado ciò, il lavoro svolto può essere visto in modo positivo anche come punto di partenza per una futura analisi ancora più approfondita.

Bibliografia

- U. Sòstero, P. Ferrarese, Analisi di Bilancio. Strutture formali, indicatori e rendiconto finanziario, Giuffrè Editore, 2000.

- E. J. Elton, M. J. Gruber, S. J. Brown, W. N. Goetzmann, Modern Portfolio Theory and Investment Analysis, Wiley, 9th Edition, 2014.

- F. Betti, Value at Risk: La gestione dei rischi finanziari e la creazione del valore, Milano, Il Sole 24 Ore, 2001.

- B. Everitt, T. Hothorn, An introduction to applied multivariate analysis with R, Springer, 2011.

- D. T. Christopoulos, Introducing Unit Invariant Knee (UIK) as an objective choice for elbow point in multivariate data analysis techniques, 2016.

- R. Johnson, D. Wichern, Applied Multivariate Statistical Analysis, Pearson Education, 6th Edition, 2014.

Ringraziamenti

A conclusione di questo elaborato desidero ringraziare tutte le persone che mi hanno accompagnato durante il percorso di laurea triennale, in particolare tutta la mia famiglia e Chiara.
Inoltre, ringrazio il mio relatore Massimiliano Caporin per avermi seguito durante la stesura di questa tesi, e la città di Padova dove ho avuto la fortuna di vivere negli ultimi tre anni.