POLITECNICO DI TORINO

Corso di Laurea in Ingegneria Informatica

Tesi di Laurea Magistrale

Clustering of historical stock price series

Relatori
prof.ssa Elena Maria Baralis
prof. Luca Cagliero

Candidato
Enrico Giuseppe Grasso

Anno accademico 2019-2020

Cos'è più importante, la connettività mondiale o il vaccino per la malaria? [...] Amo ancora le vicende IT, ma se vogliamo migliorare le nostre vite dobbiamo occuparci di questioni ben più elementari come la sopravvivenza dei bambini e le risorse alimentari.

(William Henry Gates III)

Indice

Capitolo 1

Introduzione

1.1 Contesto

Negli ultimi anni l'analisi del dato è diventato un ambito di ricerca sempre più importante. Questo principalmente perché si è registrata una notevole crescita delle fonti, degli strumenti e delle tecniche per generare dati. Questa crescita ha fatto si che si immagazzinassero una enorme mole di dati con il quale però le tecniche di analisi tradizionali si sono rivelate inefficienti e limitanti. È stato necessario dunque definire nuove strategie che estraessero le informazioni utili e le rendessero fruibili. Il processo che permette di farlo prende il nome di "*Knowledge Discovery from Data*" (KDD) ed inizia l'analisi dai dati grezzi che vengono poi trasformati e a cui vengono applicati algoritmi di data mining. Questi algoritmi consentono l'estrazione delle informazioni utili dai nostri dati.

Gli algoritmi di data mining sono molteplici e vengono applicati in tantissimi ambiti. All'interno di questo lavoro verranno applicati su un dataset formato da serie temporali che descrivono l'andamento del prezzo dei titoli azionari dell'indice Standard & Poor 500, di cui fanno parte le 500 aziende statunitensi a maggiore capitalizzazione. L'obiettivo del lavoro è applicare a questo dataset un algoritmo di clustering, una fra le tecniche di analisi più diffuse. Un algoritmo di clustering ha l'obiettivo di dividere i dati in gruppi omogenei (cioè contenenti oggetti simili tra loro) e/o identificare possibili outliers (oggetti considerati diversi da tutti gli altri). Per valutare la similarità tra oggetti è necessario definire una misura di distanza tra essi.

Il clustering di serie temporali che descrivono l'andamento dei prezzi dei titoli è un utile strumento a supporto della gestione degli investimenti. L'individuazione delle correlazioni tra titoli è infatti un principio base per la diversificazione del portafoglio. L'obiettivo della diversificazione è diminuire il rischio degli investimenti tramite la presenza in portafoglio di più attività finanziarie il cui andamento non è correlato. In questo modo, ad esempio, nel caso in cui un titolo abbia performance negative l'impatto sul portafoglio viene mitigato dalla presenza di altri titoli non correlati. Fin dagli anni 60' si è tentato di studiare queste correlazioni tra titoli con un approccio basato sulla classificazione, ad esempio, per regione geografica, settore industriale o capitalizzazione. Oggi queste classificazioni vengono considerate inadatte a gestire correttamente le strategie di diversificazione. Risulta necessario quindi ridefinire i titoli considerati simili. Attorno a questo tema si è creato quindi interesse sia in ambito accademico, per ricercare quali possano essere le migliori tecniche di clustering per questo genere di serie temporali, sia in ambito finanziario per utilizzare queste tecniche.

1.2 Problematica da affrontare

Il clustering di serie temporali presenta alcune sfide nuove rispetto al clustering tradizionale. La principale difficoltà è legata alla definizione della misura di similarità. Negli ultimi anni la ricerca in questo campo ha generato un numero sempre crescente di misure di similarità adatte alle serie

temporali. Questo perché gli esperimenti empirici hanno dimostrato che la scelta di una misura appropriata è strettamente legata al dominio applicativo in cui si utilizza. Spesso la misura della distanza si scontra nei datasets reali con problemi di rumore, scala, traslazione o discontinuità. Questi fattori rendono la definizione della misura di similarità la più grande sfida nel clustering di serie temporali.

È anche importante valutare quanto debbano essere lunghe le serie temporali da analizzare. Questa tematica deve essere affrontata sia da un punto di vista applicativo (quali sono i requisiti della mia analisi?) sia da un punto di vista tecnico (quali lunghezze mi permettono di ottenere un buon clustering?).

Il clustering di serie temporali che descrivono l'andamento del prezzo dei titoli, per come lo vogliamo interpretare, presenta inoltre alcune particolarità rispetto al clustering di serie temporali tradizionale:

- è possibile analizzare le serie in formato nativo, mantenendo la totalità dell'informazione originale;
- non è necessario applicare trasformazioni arbitrarie che possono perturbare la qualità delle analisi successive;
- è possibile considerare la sequenza temporale dei dati in modo esplicito, senza l'uso di trasformazione che rendano tale correlazione implicita.

Al fine di creare uno strumento utile alla gestione del portafoglio, a differenza del generico clustering di serie temporali, è un prerequisito fornire delle misure quantitative della somiglianza, tra singoli titoli o tra titoli appartenenti ad un cluster, in modo da dare un'indicazione misurabile all'analista. Inoltre risulta necessario anche poter settare i parametri che regolano la formazione dei cluster in modo che, differenti esecuzioni con differenti parametri, producano risultati diversi da poter confrontare.

1.3 Obiettivo della tesi

Questo lavoro propone l'utilizzo dell'algoritmo di clustering K-Shape e l'implementazione di una metodologia che consente di definire i parametri di similarità. La scelta di K-Shape, presentato all'interno del paper dal titolo "*K-Shape Efficient and Accurate Clustering of Time Series*" di *John Paparrizos* e *Luis Gravano*[1] è motivata principalmente dagli ottimi risultati che ha registrato in termini di accuratezza ed efficienza indipendentemente dal dominio in cui è stato applicato.

L'obiettivo che ci si pone in questo lavoro è duplice:

1. individuare i cluster all'interno del nostro dataset;
2. individuare i singoli titoli con andamento simile tra loro.

Inoltre ci si pone anche l'obiettivo di determinare dei parametri quantitativi che descrivano gli scenari analizzati. In particolare si vuole: (a) definire una misura di similarità intra-cluster al fine di quantificare quanto i titoli appartenenti ad un cluster siano simili tra loro; (b) definire una misura di similarità tra singoli titoli: al fine di quantificare quanto due titoli siano simili tra loro.

Al fine di verificare sperimentalmente le tecniche e le metodologie proposte ci si pone l'obiettivo inoltre di applicarle sperimentalmente su dati reali.

1.4 Metodologie applicate

Il K-Shape ha anche due svantaggi per la nostra applicazione. Per prima cosa non è deterministico: si basa infatti su un'assegnazione iniziale delle serie ai cluster completamente randomica. Partendo

quindi dallo stesso input potrebbe darci in output dei cluster differenti. Inoltre l'algoritmo assegna tutte le serie ad un cluster quindi non contempla la possibilità che un oggetto possa non fare parte di nessun cluster. Per ovviare a questi svantaggi si è definita una metodologia che prevede di eseguire l'algoritmo n volte, con n definito come parametro di input. Da queste esecuzioni viene creata una matrice, detta "matrice di similarità". La matrice di similarità ha i simboli che identificano gli oggetti (le singole serie temporali) in ascissa e in ordinata. Nelle celle all'interno della matrice avremo il numero di volte (compreso tra 0 e n) che, nelle n esecuzioni, i due oggetti sono stati assegnati allo stesso cluster. Questo parametro verrà chiamato "somiglianza".

In questo modo è diventato possibile:

- definire i cluster con più robustezza considerando nella definizione dei cluster quante volte, nelle n esecuzioni, degli oggetti venivano assegnati allo stesso cluster;
- escludere dall'analisi gli oggetti che, per quella fascia temporale, non possono considerarsi parte di un cluster.

1.5 Principali risultati

Per una completa comprensione dei risultati emersi si è utilizzato come riferimento l'appartenenza dei titoli ai settori industriali definiti dallo standard GISC (*Global Industry Classification Standard*) che è un tipo di tassonomia industriale che divide le aziende in gruppi industriali in base a simili prodotti e processi produttivi. È importante sottolineare che la tassonomia GISC è stata utilizzata non come ground truth ma solo come mezzo di interpretazione dei risultati ottenuti.

Si sono analizzati i dati con diversi parametri di input non solo per far emergere le differenze, ma anche per evidenziare le costanti. Elemento comune a tutte le analisi effettuate è la presenza di 4 clusters, ad elevata similarità intra-cluster, che, in accordo con lo standard GISC, è possibile chiamare: "Financials", "Utilities", "Energy Oil & Gas" e "Real Estate". Questi cluster contano in totale oltre 100 titoli sui 500 analizzati ed emergono da qualsiasi analisi anche con range temporali differenti.

Oltre al clustering si è reputato interessante dare spazio anche all'analisi della similarità tra singoli titoli. Da questa tipologia di analisi si è riscontrato, ad esempio, che i titoli più simili ad Apple (AAPL) siano di aziende produttori di semiconduttori. Oppure come le aziende più vicine ad American Airlines (AAL) siano, dopo le altre compagnie aeree, quelle che gravitano attorno al mondo dei viaggi. Leggendo i risultati è possibile fare tantissimi esempi di questo tipo.

Grazie alla tecnica di definizione dei cluster a partire dalla matrice di similarità è stato anche possibile fare emergere i diversi livelli di relazione tra titoli di uno stesso GISC Sector. Ad esempio, fra tutti i titoli del settore "Helth Care", sono stati individuati diversi sub-cluster a cui è stato possibile dare un nome grazie alla definizione delle "GISC Industries" della tassonomia GISC. I risultati di questo tipo di analisi evidenziano anche outier o titoli molto simili che appartengono a Industries differenti. La visualizzazione tabellare di questi dati spesso non è di facile comprensione, per questo motivo è stata ideato e proposto un nuovo metodo di visualization che renda i risultati chiari e fruibili da chiunque.

Capitolo 2

Analisi dello stato dell'arte

2.1 Clustering di time-series

La profiliferazione dei dati temporali in molte discipline ha generato un sostanziale interesse nell'analisi e nel mining delle serie storiche. In questa sezione vedremo quali sono le principali sfide e le principali componenti del clustering di questi dati. All'interno di questo lavoro si utilizzeranno per indicare una time-series indistintamente anche i termini "serie temporale" o "serie storica". Per time-series si intende una successione di dati osservati su un determinato fenomeno ordinati secondo la variabile tempo. Questa sequenza numerica descrive quindi il cambiamento di un valore in funzione del tempo in cui ogni punto è il risultato di un'osservazione. Il campionamento non avviene necessariamente con cadenza uniforme. Ogni serie può essere osservata sia come un unico oggetto nella sua interezza sia come un sottoinsieme di punti. È da considerare che, a seconda delle applicazioni, una serie storica può avere anche dimensionalità molto alta.

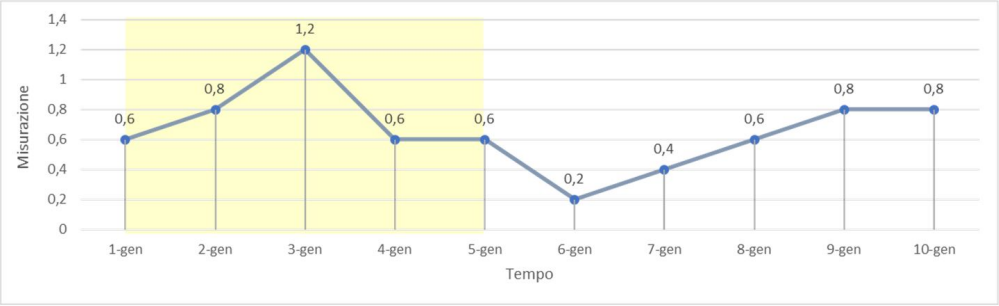

Figura 2.1. Esempio di serie temporale con campionamento giornaliero

Nella serie temporale d'esempio in Figura 2.1 abbiamo il risultato di alcune misurazioni avvenute con cadenza uniforme giornaliera dall'1 al 10 gennaio. Possiamo considerare questa serie come un unico oggetto oppure considerarne solo una parte (ad esempio solo la parte evidenziata in giallo).

Definizione 1: Clustering di Serie Temporali[2] Dato un dataset di n time-series $D = \{F_1, F_2, ..., F_n\}$ viene chiamato Clustering di Serie Temporali il processo di unsupervised partitioning di D in $C = \{C_1, C_2, ..., C_k\}$ in modo da formare gruppi omogenei di time-series sulla base di una certa misura di somiglianza.

C_k viene chiamato "Cluster", dove $D = U_{i-1}^{k} C_i$ e $C_i \cap C_j = \emptyset$ per $i \neq j$.

Il clustering è uno dei più popolari metodi di data mining, non solo grazie al suo potere esplorativo, ma anche come step di preprocessing o subroutine di altre tecniche. La maggior parte delle applicazioni del clustering di serie temporali possono essere classificate in 2 categorie:

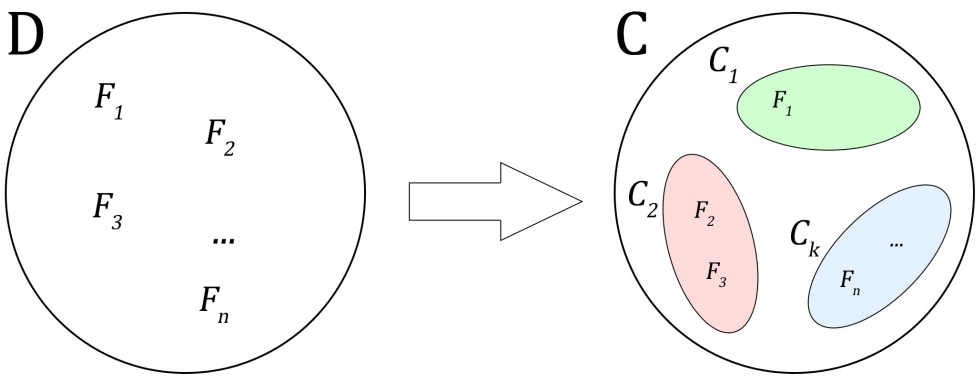

Figura 2.2. Esempio di Clustering

1. **Whole time-series clustering**: è il clustering di un gruppo di serie temporali considerate per intero e singolarmente. Viene calcolato quindi in base alla similarità tra intere serie;
2. **Subsequence clustering**: è il clustering di un set di sotto-sequenze di serie temporali che sono estratte tramite sliding window. Questo genere di clustering può essere anche utilizzato sulle sotto-sequenze all'interno di una stessa serie. La similarità viene valutata in questo caso tramite il confronto tra le sotto-sequenze.

In questo lavoro ci concentreremo sul "Whole time series clustering". Infatti lo scopo sarà quello di confrontare l'andamento di serie numeriche al fine di identificare quali hanno comportamenti simili in uno stesso range temporale. Non sarà oggetto di analisi il confronto in tempi differenti (e quindi non verranno utilizzate sliding window).

Non è scontato che il clustering debba essere utilizzato solo per individuare pattern frequenti all'interno di un dataset. È possibile infatti anche individuare dei singoli elementi che si comportano diversamente rispetto alla totalità o quasi totalità del dataset.

Le applicazioni in cui il mondo accademico sta studiando il clustering di serie temporali sono molteplici e a volte anche molto distanti tra loro. Alcuni ambiti sono: Astronomia[3], Biologia[4], Clima[5], Energia[6], Finanza[7][8][9], Medicina[10], Psicologia[11], Robotica[12], Speech/voice recognition[13], User analysis[14].

2.2 Whole Time Series Clustering

Come detto questo tipo di clustering considera le serie numeriche nella loro interezza. Questo approccio si fonda generalmente su quattro punti chiave che in questo paragrafo andremo ad esaminare:

- Representation method;
- Similarity/dissimilarity measure;
- Prototypes;
- Cluster Algorithm.

2.2.1 Representation method

La prima componente esplora come rappresentare i nostri dati prima di sottoporli all'algoritmo di clustering. È possibile utilizzare le tecniche tradizionali che mirano a migliorare la qualità del dato. Per le serie temporali però spesso abbiamo a che fare con serie temporali ad alta dimensionalità quindi per questo motivo un ruolo chiave dei representation method è svolto dalle funzioni di "dimensionality reduction". Queste tecniche consistono nel rappresentare la serie numerica grezza in un nuovo spazio dimensionale più piccolo tramite un processo che ne estrae le caratteristiche salienti. Questo processo è molto importante perché la main memory potrebbe essere non sufficiente per contenere tutti i dati (cosa spesso possibile con le serie numeriche grezze) e quindi saremmo obbligati a fare le nostre operazioni utilizzando spesso la memoria disco con conseguente aumento di tempo di calcolo che a volte potrebbero diventare insostenibili. In ogni caso la riduzione dimensionale migliora significativamente anche le performance dell'algoritmo di clustering, a prescindere dalla memoria. È anche importante in questa fase occuparsi della distorsione e del rumore presente nei nostri dati perché l'utilizzo delle serie numeriche grezze potrebbe generare cluster simili per il rumore piuttosto che per il loro comportamento e la loro forma. Viste le considerazioni fatte la scelta di un metodo opportuno di riduzione dimensionale è un elemento chiave dell'algoritmo in quanto alta dimensionalità e rumore sono caratteristiche di molte serie numeriche. Il pre-processing ci consente quindi sia di migliorare l'efficienza sia l'accuratezza della soluzione. Su questo argomento in letteratura si trova uno studio specifico [15] in cui vengono comparati 8 metodi di rappresentazione su 38 datasets. Vengono comparati alcuni metodi in cui è possibile definire, a seconda dell'applicazione, una differente compression ratio ed altri invece in cui il grado di compressione è definito automaticamente in base alle time-series. È importante sottolineare come la maggior parte degli studi fatti si basi su datasets di esempio e non su dati reali. Resta quindi fondamentale studiare il representation method opportuno in base alla specifica applicazione.

2.2.2 Similarity/dissimilarity measure

Il clustering di serie temporali presenta alcune sfide nuove rispetto al clustering tradizionale. La principale difficoltà è legata alla definizione della misura di similarità. La ricerca di una misura della distanza tra serie temporali è stata teorizzata per la prima volta nel 1993[16]. Negli ultimi anni la ricerca in questo campo ha generato un numero sempre crescente di misure di similarità adatte alle serie temporali. Questo perché gli esperimenti empirici hanno dimostrato che la scelta di una misura appropriata è strettamente legata al dominio applicativo in cui si utilizza. Essendo questa scelta determinante per il successo dell'algoritmo di clustering sono state proposte quindi diverse soluzioni. A seconda dell'applicazione possiamo distinguere tre differenti tipologie di misura della similarità per il clustering di time-series:

1. **Shape-based distances**: per comparare due serie si cerca di rendere le loro forme quanto più simili possibili tramite contrazioni e allungamenti non lineari. Possiamo definire due sottocategorie di misure: (a) Lock-step measure in cui entrambe le serie hanno uguale

lunghezza (n=m) ed ogni punto i della serie x viene comparato con il corrispettivo punto i della serie y; (b) Elastic measure in cui il calcolo è più flessibile e consente comparazioni tra punti anche one-to-many;

2. **Feature-based distances**: dalle serie temporali viene creato un feature vector di dimensione minore. Solitamente tutte le serie, a prescindere dalla loro dimensione iniziale, vengono convertite in un feature vector di uguale dimensione. Dopo viene calcolata la distanza tra i feature vector e viene applicato un algoritmo di clustering tradizionale utilizzando, ad esempio, la distanza Euclidea. Gli approcci feature-based vengono usati spesso anche per ottenere riduzione della dimensionalità e del rumore;

3. **Model-based distances**: ogni serie viene convertita in un modello parametrico. Questo approccio presenta però problemi di scalabilità.

Cerchiamo adesso di partire dalla definizione più semplice di misura di distanza per serie temporali.

Definizione 2: Univariate time-series Viene definita univariate time-series una serie temporale formata da una sequenza di numeri reali campionati ad intervalli regolari. Il metodo più semplice per calcolare la distanza tra due serie temporali è considerarle come univariate time-series e dopo calcolare la distanza misurandola tra tutti i punti. La distanza tra tutti i punti di due serie temporali viene definita come la somma delle distanze tra i singoli punti. Cioè dato un dataset di n time-series $D = \{F_1, F_2, ..., F_n\}$ la distanza tra la serie F_i e la serie F_j viene definita come $dist(F_i; F_j) = \sum_{t=1}^{T} dist(f_{it}, f_{jt})$ dove t indica il tempo della misurazione.

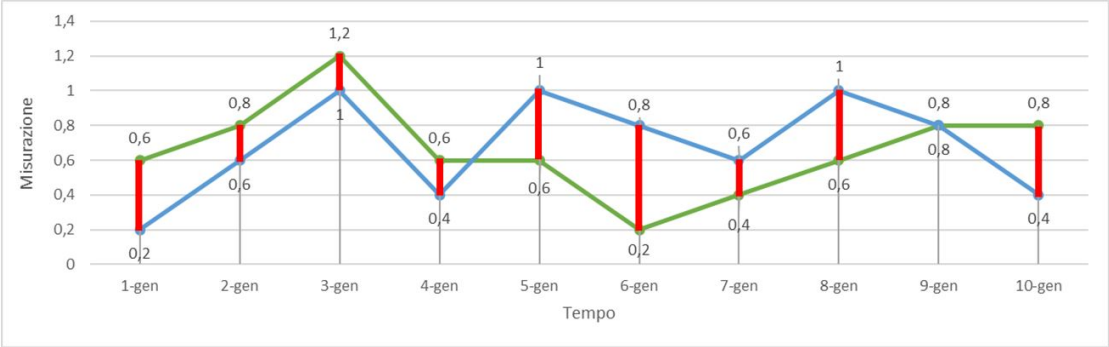

Figura 2.3. Esempio di misura della distanza tra serie temporali

In Figura 2.3 vediamo in verde la serie F_1 e in azzurro la serie F_2 . Le linee rosse indicano un possibile calcolo della distanza tra i singoli punti delle due serie la cui somma sarà data da $\sum_{t=1}^{T} dist(f_{it}, f_{jt})$ Nel corso degli anni però sono state definite parecchie misure più complesse rispetto a quella appena descritta. Alcune sono specifiche per un representation method, altre sono indipendenti e possono essere applicate anche direttamente ai dati grezzi. Una differenza sostanziale tra il clustering tradizionale e il clustering di serie temporali è che nel primo caso la distanza tra due oggetti viene definita in modo preciso, mentre nel secondo viene spesso calcolata per approssimazione.

Le misure di distanza più utilizzate sono:

- **Distanza di Hausdorff**[17]: la distanza di Hausdorff tra due serie temporali A e B è il più piccolo numero positivo r, per il quale qualsiasi punto di A è entro la distanza r da qualsiasi punto di B, e qualsiasi punto di B è entro la distanza r da qualsiasi punto di A;

- **Modified Hausdorff (MODH)**[18]:diverse varianti della "Distanza di Hausdorff";

- **Hidden Markov Models based distance (HMM)**[19]: viene utilizzato il modello HMM per derivare in modo probabilistico una nuova misura di distanza fra sequenze. "Hidden Markov Models (HMM)" descrive un processo aleatorio in cui la probabilità del passaggio da

uno stato del sistema ad un altro dipende solo dallo stato di partenza (proprietà di Markov) e non dal percorso che ha portato a questo stato. In particolare viene definito "Hidden" quando gli stati non sono osservabili direttamente;

- **Dynamic Time Warping (DTW)**: discussa successivamente;
- **Distanza Euclidea**: discussa successivamente;
- **Distanza Euclidea in un sottospazio PCA (Principal component analysis)**: viene applicata la distanza euclidea non su tutto lo spazio ma sulle componenti più significative;
- **Longest Common Sub-Sequence (LCSS)**[20]: si basa sulla ricerca della maggiore sottosequenza comune ad un set di sequenze (spesso due).

La scelta della misura della distanza appropriata dipende dalle caratteristiche delle serie temporali, dalla lunghezza, dal representation method e chiaramente dall'obiettivo del clustering. Spesso la misura della distanza si scontra nei datasets reali con problemi di rumore, scala, traslazione o discontinuità, proprietà che sono comuni a molte serie. Questi fattori rendono la definizione della misura di similarità la più grande sfida nel clustering di serie temporali. In generale gli approcci possibili nella ricerca della similarità sono: (a) relativi al tempo: che sono basate sul calcolo della distanza ad ogni step di tempo. Questo tipo di misurazione è molto costosa sulle serie grezze quindi viene spesso applicata dopo una trasformazione; (b) relative alla forma: queste misure sono calcolate indipendentemente dalla distanza tra i punti. A livello accademico la similarità nel tempo viene considerata un caso particolare della similarità nella forma; (c) similarità strutturale: in queste misure la similarità è basata sui parametri con cui le time-series vengono modellate. Questo approccio è adatto per serie lunghe. Non è semplice scegliere la giusta misura di distanza. Non esiste in letteratura una metodologia formale per scegliere qual è la misura più appropriata per i nostri dati. Con l'obiettivo di semplificare questo task è stato proposto[21] un framework di classificazione che fornisce un metodo per selezionare automaticamente le misure di distanza più adatte per il clustering di un dataset di serie temporali. Questo metodo estrae le caratteristiche principali di un dataset di time series e propone automaticamente la migliore misura di distanza da un set di possibili candidati.

Vediamo adesso brevemente le due misure più utilizzate nel clustering di time series: la distanza Euclidea e la Dynamic Time Warping (DTW).

Euclidean Distance

Date due time-series T ed S, formate da N misurazioni in cui $T = \{T_1, T_2, ..., T_n\}$ e $S = \{S_1, S_2, ..., S_n\}$ si definisce Distanza Euclidea tra T ed S :

$$ED(T, S) = \sqrt{\sum_{n=1}^{N} (r_n - s_n)^2}$$

Nonostante la sua semplicità la distanza Euclidea è competitiva in molte applicazioni[15]. È importante però sottolineare che è necessario rimuovere, prima di applicare la distanza Euclidea, le possibili distorsioni o il rumore presente sui dati grezzi. Se non è possibile dovranno essere utilizzate misure di distanza diverse. Inoltre, questa misura, si presta ad essere utilizzata solo per la comparazione di serie di uguale lunghezza e uguale campionamento al fine di andare a comparare punto per punto misurazioni effettuate nello stesso momento.

In Figura 2.4 vediamo un esempio di come viene calcolata la distanza tra due serie T ed S.

Dynamic Time Warping (DTW)

La Dynamic Time Warping (DTW) viene introdotta per superare i limiti della distanza euclidea [23]. Viene utilizzata infatti per la maggior parte per il confronto di serie non sincronizzate tra

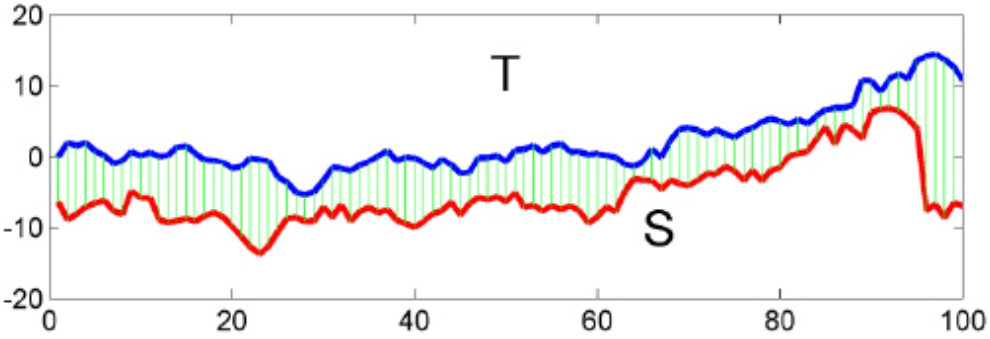

Figura 2.4. Euclidean distance su due serie numeriche T ed S[22]

loro, cioè serie in cui punti simili potrebbero essere in posizioni differenti nel tempo. Con la distanza euclidea non abbiamo la possibilità di vedere questo genere di somiglianza.

Per determinare la distanza DTW tra due serie temporali T ed S viene calcolata una matrice dei costi (Local Cost Matrix - LCM) di dimensione $n \times m$ dove ogni elemento $(i;j)$ rappresenta la distanza tra s_i e t_i. Questa distanza viene solitamente definita come differenza quadratica $d(s_i;t_i) = (s_i - t_i)^2$. Successivamente viene definito un percorso (warping path) $W = \{w_1, w_2, ..., w_K\}$ dove $max(n,m) \leq K \leq m+n-1$. Questo percorso deve attraversare la LCM rispettando tre condizioni:

1. **Boundary condition**: il percorso deve iniziare e finire agli angoli della matrice: $w_1 = (1;1)$ e $w_k = (n;m)$
2. **Continuity**: solo gli elementi adiacenti possono essere considerati come step successivi del percorso (incluse le diagonali). Quindi se $w_q = (i;j)$ allora l'elemento $w_{(q+1)}$ è uno tra: $(i+1;j)$,$(i;j+1)$ oppure $(i+1;j+1)$ con $q = 1, ..., K-1$, $i = 1, ..., n-1$ e $j = 1, ..., m-1$.
3. **Monotonicity**: gli step successivi del percorso devono essere spazialmente monotoni nel tempo.

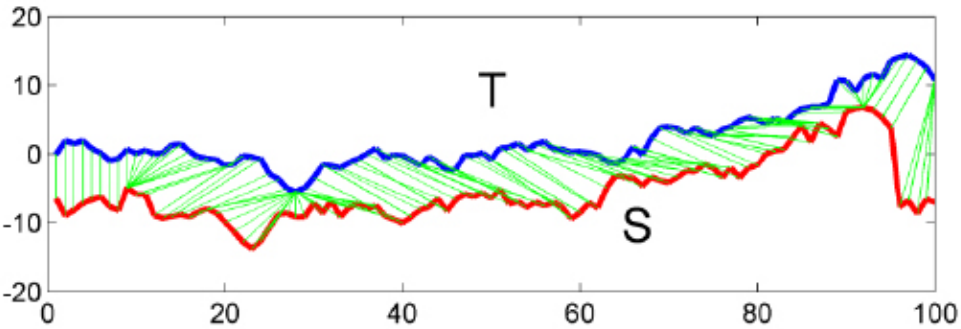

Figura 2.5. Dynamic Time Warping su due serie numeriche T ed S[22]

La distanza totale per il percorso W è ottenuta dalla somma degli elementi della matrice LCM che sono toccati dal Warping Path. La distanza Dynamic Time Warping è derivata dal percorso con la minima distanza totale. In letteratura ci sono diversi modi per definire questa distanza. Vediamo quella che viene calcolata con la radice della somma[24]:

$$d_D(t,s) = min\sqrt{\sum_{k=1}^{K} w_k}$$

Da notare che questa distanza è uguale alla distanza euclidea con $n = m$ e con il warping path definito attraversando diagonalmente la matrice LCM. In Figura 2.5 vediamo come cambiano i punti scelti per il calcolo della distanza tra le due serie T ed S rispetto ai punti scelti per la distanza Euclidea che vediamo in Figura 2.4.

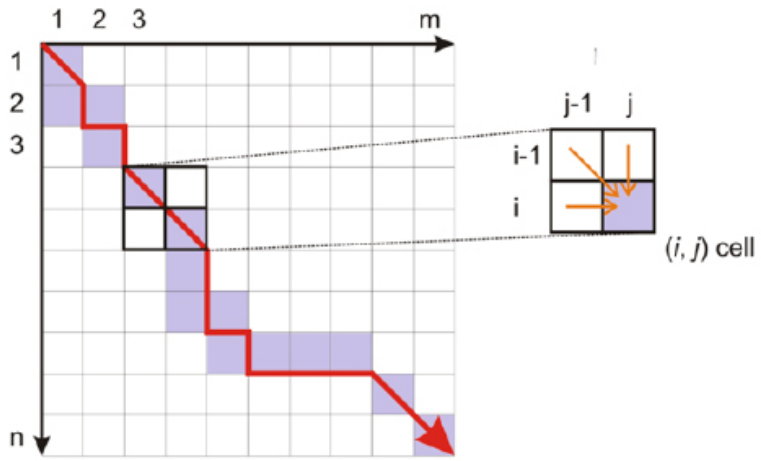

Figura 2.6. Warping path sulla Local Cost Matrix (LCM)[22]

2.2.3 Prototypes

Alcuni algoritmi di clustering hanno la necessità di poter definire ogni cluster con un cluster prototype o centro del cluster cioè un oggetto che è rappresentativo di tutti gli oggetti del cluster. L'obiettivo è quello di individuare un prototype che rappresenti al meglio l'intero cluster. Questa ricerca è essenzialmente una subroutine del clustering. Date le time-series di un cluster si definisce il Prototype R_j del cluster la serie che minimizza la distanza tra tutte le time-series del cluster e lo stesso Prototype R_j. Sono stati proposti diversi metodi per calcolare il cluster prototype di serie temporali ma molte delle pubblicazioni al riguardo definiscono il metodo ma non provano la sua correttezza. In ogni caso si possono definire 3 approcci per il calcolo del prototype:

1. **Medoid**: è il metodo più comune e consiste nell'utilizzare il cluster medoid come prototype. Il cluster medoid altro non è che la serie temporale più rappresentativa fra un gruppo di serie temporali (nel nostro caso un cluster). Viene definita più rappresentativa la serie all'interno del cluster che minimizza la somma delle distanze al quadrato rispetto alle altre serie del cluster stesso. Dato un cluster viene calcolata la distanza fra tutte le coppie di time-series utilizzando ad esempio la distanza Euclidea o DTW. Successivamente la time-series del cluster che ha la più bassa somma al quadrato delle distanze è definita medoid del cluster.

2. **Average**: se le serie sono di uguale lunghezza allora può essere definito come prototype la time-series in cui ogni punto è pari alla media tra tutte le time-serie.

3. **Local search**: questo approccio è un mix tra l'approccio medoid e l'average.

2.2.4 Clustering

Le tecniche di clustering di serie temporali oggi esistenti sono molteplici. In generale alcune sono usate sui dati grezzi, altre su dataset che hanno già subito una dimension reduction. Generalmente il clustering di serie temporali può essere suddiviso in:

1. **Hierarchical clustering** È un approccio di cluster analysis che crea una gerarchia di clusters usando algoritmi agglomerativi o divisivi. Gli algoritmi agglomerativi partono considerando ogni elemento come un cluster singolo e dopo gradualmente li uniscono tra loro (approccio bottom-up). Gli algoritmi divisivi al contrario partono con tutti gli oggetti in un unico cluster e successivamente lo frazionano fino ad avere tutti i clusters con un solo elementi (approccio top-down). In generale l'approccio gerarchico non è molto solido perché non può correggersi dopo uno split o dopo un merge dei clusters. In molti studi gli algoritmi gerarchici vengono utilizzati invece per valutare una dimensional reduction o una misura di distanza. In contrasto con molti algoritmi invece questo metodo non richiede la definizione a priori del numero di clusters. Questo punto è molto interessante perché nelle serie numeriche spesso è difficile definire il numero di clusters. Inoltre, il clustering gerarchico, permette anche di clusterizzare serie con differente lunghezza se viene utilizzata un'appropriata misura di distanza tra time-series. Può essere utile per piccoli dataset in quanto non è molto scalabile.

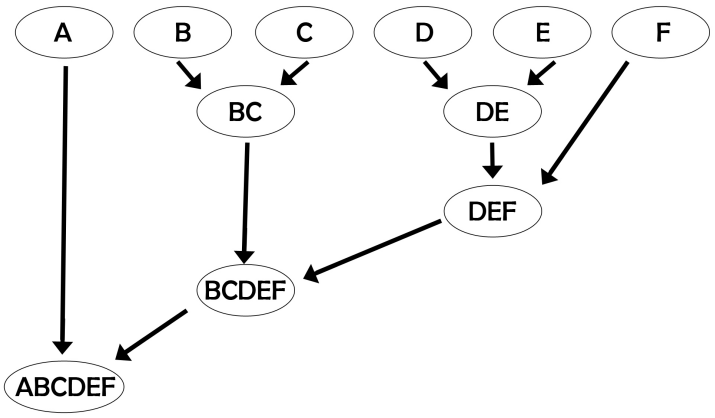

Figura 2.7. Dendogramma di esempio per algoritmi di clustering gerarchici

2. **Partitioning clustering** Gli algoritmi di partitioning clustering creano k gruppi da n oggetti unlabelled in modo che ogni gruppo contenga almeno un oggetto. Uno dei più usati algoritmi di questo tipo è il K-Means[25] dove ogni cluster ha un prototype (centroid) che è il valore medio dei suoi oggetti. L'idea dietro il K-Means è la minimizzazione della distanza totale (tipicamente la distanza Euclidea) fra tutti gli oggetti di un cluster dal loro centro del cluster (prototype). Un altro esempio di algoritmo di partitioning è il k-Medoids (PAM)[26] dove il prototype di ogni cluster è l'oggetto più vicino al centro del cluster. Sia per k-Means che per k-Medoids è necessario definire a priori il numero di cluster. Questo li rende non utilizzabili in molte applicazioni. Tuttavia questi metodi sono molto più efficienti in termini computazionali rispetto agli algoritmi gerarchici e numerosi studi testimoniano che sono adatti per il clustering di time-series. k-Means e k-Medoids sono tecniche in cui ogni oggetto o appartiene ad un cluster o non ci appartiene. Esistono algoritmi corrispettivi, FCM (Fuzzy c-Means)[27] e Fuzzy c-Medoids[28] che creano dei "soft" cluster in cui ogni membro ha un grado di appartenenza al cluster.

3. **Model-based clustering** In questo approccio si presume che i dati siano generati da una combinazione di distribuzioni di probabilità. Con questa tipologia di clustering si tenta di recuperare il modello originale cercando di ottimizzare l'adattamento dei dati al modello. Ciascun modello recuperato dai dati generati definisce un cluster diverso. In generale questo approccio ha due svantaggi: in primo luogo ha bisogno di alcuni parametri iniziali che sono basati su assunzioni che potrebbero essere errate e quindi generare risultati errati; in secondo luogo è molto lento specialmente in grandi dataset.

4. **Density-based clustering** Questo metodo si basa sull'intuizione di individuare i cluster come sottospazi in cui ci sia una maggiore densità di oggetti separati da sottospazi in cui ci sia densità minore di oggetti. Uno dei principali algoritmi che lavora con un approccio

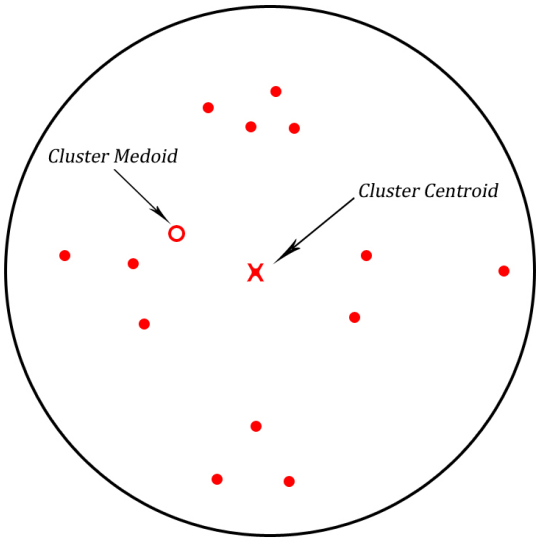

Figura 2.8. Esempio di cluster con evidenziati Medoid e Centroid

density-based è DB-Scan [29]. In questo algoritmo si considera un punto q raggiungibile da un altro punto p se la loro distanza è minore di un valore assegnato ε. I punti che saranno tra loro raggiungibili vengono considerati parte dello stesso cluster. In la letteratura questo approccio non è stato molto utilizzato per il clustering di serie temporali a causa della sua complessità.

2.3 Clustering evaluation

In questa sezione esploriamo i metodi di valutazione degli algoritmi di clustering di serie temporali. Su questo ambito esiste uno studio composto da diversi articoli[23] che conclude che la valutazione del mining di serie temporali deve seguire alcune regole fondamentali tra cui:

- La validazione dell'algoritmo deve avvenire su un ampio range di dataset a meno che non sia specifico di un'applicazione. I dataset devono essere pubblicati e disponibili gratuitamente;
- L'implementation bias deve essere evitato con un'attenta progettazione degli esperimenti;
- Se possibile i dati e gli algoritmi devono essere resi disponibili gratuitamente;
- Nuovi metodi di misurazione della distanza devono essere comparati con metriche semplici e stabili come la distanza Euclidea.

In generale la valutazione di un algoritmo può essere complessa in assenza di dati oggettivi come le data labels. La definizione del cluster dipende dall'utente, dal dominio ed è una cosa soggettiva. Per esempio il numero di cluster, la loro dimensione, la definizione di outlier o la definizione di similarità tra serie temporali sono tutti concetti che dipendono dall'obiettivo del mio lavoro e devono essere dichiarati a priori. In ogni caso, definite le label per i nostri dati o tramite la decisione umana o a partire dalla loro generazione (se sono dati sintetici) i risultati devono poter essere valutati tramite qualche misurazione. Le label definite da un umano non saranno perfette in termini di clustering ma in pratica ci aiuteranno a capire quali siano i punti di forza e di debolezza del nostro algoritmo. Generalmente per la valutazione dell'algoritmo viene generato un singolo numero reale che rappresenterà l'accuratezza dei differenti algoritmi di clustering. Le misure che sono utilizzate per valutare gli algoritmi vengono divise in due categorie:

1. **External index**: questo tipo di valutazione genera un indice che viene utilizzato per misurare la similarità tra due clustering degli stessi dati: il primo formato dall'algoritmo in esame, il secondo quello considerato come ground truth. Si definisce "ground truth" il clustering ideale dei nostri dati, spesso generato da umani esperti del settore. Questo è il metodo di valutazione più usato.

   Uno degli external index è il "Cluster purity". Considerando $G = \{G_1, G_2, ..., G_M\}$ come i cluster ground truth e $C = \{C_1, C_2, ..., C_M\}$ come i cluster generati dall'algoritmo da valutare. Considerando che ad ogni cluster di C viene assegnata la label più frequente all'interno del cluster la purity dei cluster C rispetto ai cluster G è pari al numero di elementi correttamente assegnati ai cluster diviso il numero totale degli elementi. Un buon algoritmo di clustering avrà purity vicino ad 1 mentre un algoritmo scadente avrà purity vicina allo 0.

2. **Internal index**: questo tipo di valutazione viene realizzata invece non considerando nessuna informazione esterna all'algoritmo. Viene formalizzato l'obiettivo di avere un'alta somiglianza intra-cluster (gli elementi di uno stesso cluster saranno simili) e una bassa somiglianza inter-cluster (elementi di cluster differenti saranno dissimili). Questo approccio viene utilizzato quando la ground truth è sconosciuta. Questo genere di indice può essere utilizzato solo per comparare differenti approcci di clustering generati tramite lo stesso modello.

2.4 Data mining finanziario

2.4.1 Analisi tecnica e analisi fondamentale

Come già anticipato nella parte introduttiva, in questo lavoro si andrà ad analizzare l'andamento dei titoli dell'indice S&P500 al fine di suddividere i titoli azionari in cluster e definire un fattore di somiglianza tra singoli titoli. Esiste un'ampia letteratura sullo studio dei mercati finanziari che si suddivide in due grandi categorie: l'analisi tecnica e l'analisi fondamentale. L'analisi tecnica è "lo studio del movimento del mercato, o market action, tramite l'uso sistematico di grafici, allo scopo di prevedere le tendenze future dei prezzi"[30]. In contrapposizione l'analisi fondamentale studia tutti i fattori rilevanti per determinare il giusto prezzo dell'azione. "Entrambe le analisi cercano di determinare la direzione dei prezzi da punti di vista diversi: l'analisi fondamentale studia le cause dei movimenti del mercato, mentre l'analisi tecnica ne studia gli effetti. (...) Alcuni dei più forti movimenti al rialzo o al ribasso della storia si sono avviati con piccole e impercettibili variazioni dei fondamentali e, una volta trascorso il tempo necessario per individuarli, il trend risultava già cambiato; in questo caso però, l'analista tecnico avrebbe già potuto operare nel senso giusto grazie alla sua interpretazione dei grafici. (...) Da questo si può capire come gli analisti tecnici considerano il proprio lavoro superiore rispetto agli analisti fondamentali"[30]. L'andamento dei titoli viene sicuramente influenzato dalle regole definite dall'analisi tecnica. Questo fattore viene descritto dalla "teoria di autoalimentazione" che dice che, vista la larga diffusione dell'analisi tecnica, è possibile che le ondate di acquisti o vendite potrebbero essere totalmente dipendenti da ciò che viene indicato dall'analisi tecnica dei grafici. È anche vero che l'interpretazione dei grafici è soggettiva e gli analisti non si trovano sempre d'accordo sull'analisi dei grafici quindi la possibilità che tutti gli analisti agiscano nello stesso tempo e allo stesso modo è molto remota. In contrapposizione all'analisi tecnica in ambito accademico si è affermata la "Teoria di random walk" che dice che il movimento dei prezzi è casuale ed imprevedibile e che la storia dei prezzi non costituisce un indicatore per i prezzi futuri. La teoria afferma che nei sistemi economici complessi, come le economie moderne, non vi sono delle leggi deterministiche che descrivono le fluttuazioni economiche. La ricerca di leggi scientifiche stabili che descrivano scientificamente i sistemi economici e finanziari sfortunatamente ha concluso che, sebbene esistano alcune proprietà fondamentali, esse sono variabili nel tempo in modo casuale.

In tutti i casi il clustering di titoli azionari e la definizione di misure di distanza tra essi potrebbe essere uno strumento utile sia per gli analisti tecnici sia per gli analisti fondamentali. Entrambi infatti potrebbero avere un nuovo elemento a supporto delle loro decisioni. Vediamo due possibili casi d'uso:

- **Analista tecnico**: individuato tramite grafici un pattern rialzista o ribassista per un determinato titolo, potrebbe tramite il clustering o la distanza tra i titoli identificare quali titoli potrebbero avere lo stesso comportamento;
- **Analista fondamentale**: individuata una nuova condizione (esempio: nuova regolamentazione in vigore) che potrebbe avere un effetto diretto, positivo o negativo, su un determinato titolo potrebbe tramite il clustering o la distanza tra i titoli identificare quali titoli potrebbero essere a loro volta coinvolti dalla stessa condizione.

2.4.2 Stato dell'arte su mining di dati finanziari

L'analisi dei dati finanziari è un argomento ampiamente affrontato in letteratura in molti aspetti. In questa sezione vediamo qualche esempio di articoli che trattano questo argomento. Molti di questi sono stati o verranno citati all'interno di questo lavoro.

L'articolo "*Clustering of financial time series*" [31] ad esempio affronta la tematica del clustering di serie temporali che descrivono l'andamento dei tassi di cambio dell'Euro rispetto ad altre valute internazionali. Per farlo utilizza un metodo di clustering fuzzy (dove ogni oggetto può appartenere a più cluster) basato su un modello GARCH.

In "*Clustering economic and financial time series: Exploring the existence of stable correlation conditions*"[32] invece vengono approfondite le tematiche relative al clustering di serie temporali finanziarie: le motivazioni, le definizioni di clustering e le varie tipologie di misura di distanza.

All'interno di "*Clustering Financial Time Series: How Long is Enough?*"[33] troviamo un'ampia analisi statistica basata su quanto dovrebbero essere lunghe le serie temporali finanaziarie negli algoritmi di clustering.

Il paper "*Hausdorff clustering of financial time series*"[17] esegue una procedura di clustering basata sulla distanza di Hausdorff come misura di somiglianza. Il metodo viene applicato alle serie temporali finanziarie dell'indice Dow Jones Industrial Average (DJIA).

Su "*Clustering Seasonality Patterns in the Presence of Errors*"[34] viene realizzato un algoritmo di clustering per lo studio dei pattern delle vendite stagionali che utilizza una nuova funzione di distanza che si basa sulla distribuzione di errori nei dati.

L'articolo "*Cluster financial time series for portfolio*"[7] esplora il clustering di serie temporali che descrivono l'andamento del prezzo di 100 titoli azionari ai fini di creare un portafoglio bilanciato.

2.4.3 Clustering di serie temporali finanziarie

Il clustering è divenuto molto importante come tecnica di analisi di serie temporali. Come detto non è possibile definire leggi deterministiche che prevedano l'andamento di serie temporali finanziarie. È possibile però trovare una sorta di dipendenze stabili tra le variabili economiche o le loro distribuzioni. Dal punto di vista finanziario il clustering di serie temporali può esserci utile per definire dei pattern che supportino i processi decisionali oltre che consentire una robusta gestione di risk e portafoglio ed essere di supporto agli algoritmi predittivi.

Possiamo distinguere il clustering in base a tre principali aree applicative[32]:

1. Identificazione di aree o settori per ragioni di policy-making;
2. Identificazione di similarità strutturali nei processi economici per le previsioni economiche;
3. Identificazione di stabili dipendenze per la gestione del risk e del portafoglio.

Questo lavoro si concentra nello studio di uno strumento di clustering per la gestione degli investimenti. L'individuazione delle correlazioni tra titoli è infatti un principio base per la diversificazione del portafoglio. Teorizzato all'interno del libro "Portfolio selection: efficient diversification of investments"[35] la diversificazione del portafoglio è un caposaldo della gestione degli investimenti. L'obiettivo della diversificazione è diminuire il rischio degli investmenti tramite la presenza in portafoglio di più attività finanziarie il cui andamento non è correlato. In questo modo nel caso in cui uno strumento finanziario abbia performance negative l'impatto sul portafoglio viene mitigato dalla presenza di altri strumenti non correlati. Fin dagli anni 60' si è tentato di studiare queste correlazioni tra strumenti con un approccio basato sulla classificazione, ad esempio, per regione geografica, settore industriale o capitalizzazione. Oggi queste classificazioni vengono considerate inadatte a gestire correttamente le strategie di diversificazione. Risuta necessario quindi ridefinire i settori e i titoli considerati simili. Molte aziende fin dagli inizi degli anni 2000 hanno iniziato ad utilizzare tecniche di clustering a questo scopo[32].

Capitolo 3

Algoritmo utilizzato e metodologia applicata

All'interno di questo capitolo verranno presentati le basi teoriche e le scelte implementative dell'algoritmo di clustering di serie temporali finanziarie sviluppato. Per prima cosa viene presentato l'algoritmo di clustering scelto, il K-Shape, discutendone la logica, le basi teoriche e i risultati ottenuti dai ricercatori su datasets di esempio. Successivamente viene esposta la metodologia ideata per la definizione dei clusters, per il calcolo della similarità tra singole serie temporali e per esplorare le dinamiche interne ad un cluster o ad un settore industriale. Verrà inoltre definito come calcolare un valore di similarità intra-cluster.

Come abbiamo già visto la scelta dei metodi di clustering è strettamente legata allo specifico dominio applicativo. Il clustering di serie temporali che descrivono l'andamento del prezzo dei titoli azionari, per come lo vogliamo interpretare, presenta infatti alcune particolarità rispetto al clustering di serie temporali tradizionale:

- le serie sono sempre allineate quindi non si devono analizzare porzioni di serie tramite sliding window;
- non è utile analizzare serie a dimensionalità molto alta quindi non si devono estrarre le caratteristiche salienti ma è possibile lavorare direttamente sui dati;
- le serie in ingresso devono subire necessariamente una standardizzazione quindi non vengono comparate serie temporali con uguale ordine di grandezza.

Al fine di creare uno strumento utile alla gestione del portafoglio, a differenza del clustering di serie temporali, è un prerequisito fornire delle misure quantitative della somiglianza, tra singoli titoli o tra titoli appartenenti ad un cluster, in modo da dare un'indicazione misurabile all'analista. Inoltre risulta necessario anche poter settare i parametri che regolano la formazione dei cluster in modo che, differenti esecuzioni con differenti parametri, producano risultati diversi da poter confrontare.

3.1 K-Shape

La metogologia sviluppata, come anticipato, utilizza l'algoritmo di clustering "K-Shape" presentato all'interno del paper "K-Shape Efficient and Accurate Clustering of Time Series" di John Paparrizos e Luis Gravano[1]. All'interno di questa sezione ripercorremo i punti salienti del paper citato al fine di comprendere al meglio il funzionamento dell'algoritmo.

K-Shape è un algoritmo per il clustering di serie storiche che si basa su una procedura iterativa scalabile che crea cluster omogenei e ben separati. Come già detto la misura della distanza è cruciale per qualsiasi algoritmo di clustering e, in particolare, per il clustering di serie temporali.

K-Shape usa una distanza Shape-based che è una versione standardizzata della cross correlation. Tramite questa distanza calcola i cluster centroids che poi vengono usati in ogni iterazione per aggiornare l'assegnazione della serie al cluster. All'interno del paper viene presentato K-Shape come il migliore algoritmo di clustering di serie temporali in termini di accuratezza e inoltre anche come algoritmo domain-indipendent. Per affermare ciò gli autori testano questo algoritmo su 48 dataset ottenendo risultati eccellenti.

3.1.1 Algoritmo

K-Shape è un algoritmo di Partitioning Clustering che si basa su un metodo iterativo simile al più famoso K-Means. L'algoritmo riesce a:

1. produrre cluster omogenei e ben separati;
2. avere una complessità che scala linearmente con il numero di time-series;

L'algoritmo si aspetta in input il dataset composto dalle time-series X e il numero di cluster che vogliamo produrre k. Inizialmente assegna randomicamente ogni serie che è in X a dei clusters. Poi inizia il suo ciclo di iterazioni e per ognuno esegue due step:

1. **Refinement step**: vengono ricalcolati (o calcolati la prima volta se siamo alla prima iterazione) i centroidi dei cluster per riflettere i cambiamenti dei membri dei cluster avvenuti nell'assignment step (o nell'assegnazione casuale se siamo alla prima iterazione);
2. **Assignment step**: viene aggiornata l'appartenenza al cluster comparando ogni serie con tutti i centroidi e assegnando ogni serie al cluster con il centroide più vicino;

L'algoritmo ripete questi due step fino a quando una di queste due condizioni non viene verificata:

1. Durante l'assignment step non ci sono stati cambiamenti nei membri del cluster;
2. È stato raggiunto il numero massimo consentito di iterazioni.

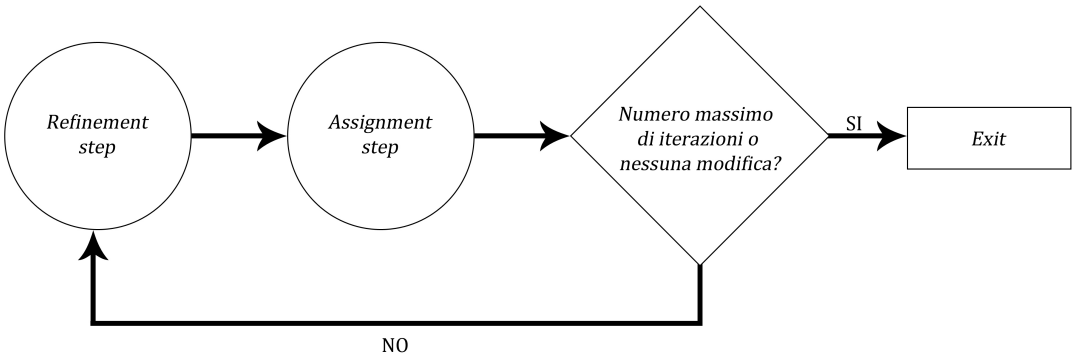

Figura 3.1. Descrizione algoritmo K-Shape

Vedremo ora i dettagli relativi alla misura della similarità tra serie e al calcolo dei centroidi. Infine analizzeremo la complessità dell'algoritmo e i risultati che ha dato su datasets di esempio.

3.1.2 Misura della similarità

K-Shape è un algoritmo che ha l'obiettivo di individuare le similarità basate sulla forma. Per raggiungere questo scopo la misura di distanza deve essere in grado di gestire distorsioni di ampiezza

e di fase. In altre parole, se trasformassimo una sequenza x in $x = ax+b$, dove a e b sono costanti, la misura della distanza fra x e altre sequenze deve essere costante. Le migliori misure di distanza, come ad esempio DTW, offrono invarianza a queste distorsioni, ma sono computazionalmente molto pesanti. Per migliorare l'efficienza in questo genere di approcci K-Shape adotta una versione standardizzata della misura di distanza "Cross-Correlation". La Cross-correlation è una misura che compara ad uno ad uno i punti tra due segnali. Non è stata mai utilizzata prima del K-Shape per la misura di distanza tra serie temporali perché il mondo accademico si è focalizzato su misure elastiche che comparano i punti one-to-many o one-to-none. È una misura statistica con il quale è possibile determinare la similarità di due sequenze $\vec{x} = (x_1, ..., x_m)$ e $\vec{y} = (y_1, ..., y_m)$ anche se non sono completamente allineate (per semplicità consideriamo sequenze di uguale lunghezza anche se la cross correlation può essere calcolata su sequenze di differente lunghezza). Per riuscire ad essere shift-invariance la cross-correlation lascia $\vec{y}$ statico e fa scorrere $\vec{x}$ su $\vec{y}$ per calcolare il loro prodotto scalare per ogni shift s di $\vec{x}$. Indichiamo lo shift di una sequenza come segue:

$$\vec{x_s} = \begin{cases} (\overbrace{0, ..., 0}^{|s|}, x_1, x_2, ..., x_{m-s}) & , s \geq 0 \\ (x_{1-s}, ..., x_{m-1}, x_m, \underbrace{0, ..., 0}_{|s|}) & , s < 0 \end{cases}$$

Dove con $\vec{x_{(s)}}$ sono considerati tutti i possibili shift con $s \in [-m, m]$. Definiamo la sequenza $CC_\omega(\vec{x}, \vec{y}) = (C_1, ..., C_\omega)$ con lunghezza $2m - 1$ definita come segue:

$$CC_\omega(\vec{x}, \vec{y}) = R_{(\omega-m)}(\vec{x}, \vec{y}), \omega \in \{1,2, ...,2m - 1\}$$

Dove $R_{(\omega-m)}(\vec{x}, \vec{y})$ è calcolato a sua volta come:

$$R_k(\vec{x}, \vec{y}) = \begin{cases} \sum_{l=1}^{m-k} x_{(l+k)} \cdot y_l & , k \geq 0 \\ R_{-k)}(\vec{x}, \vec{y}) & , k < 0 \end{cases}$$

Il nostro obiettivo è calcolare la posizione ω per il quale $CC_\omega(\vec{x}, \vec{y})$ è massimizzata. Basandoci su questo valore di ω, lo shift ottimale di $\vec{x}$ rispetto a $\vec{y}$ è quindi $\vec{x_s}$ dove $s = \omega - m$. In base al dominio applicativo possono essere richieste differenti standardizzazioni di $CC_\omega(\vec{x}, \vec{y})$. Le più comuni sono le biased estimator NCC_b, le unbiased estimator NCC_u e le coefficient normalization NCC_c che sono definite come segue:

$$NCC_q(\vec{x}, \vec{y}) = \begin{cases} \frac{CC_\omega(\vec{x}, \vec{y})}{m} & , q = "b" (NCC_b) \\ \frac{CC_\omega(\vec{x}, \vec{y})}{m-|\omega|} & , q = "u" (NCC_u) \\ \frac{CC_\omega(\vec{x}, \vec{y})}{\sqrt{R_0(\vec{x}, \vec{y}) \cdot R_0(\vec{y}, \vec{y})}} & , q = "c" (NCC_c) \end{cases}$$

Quindi la standardizzazione dei dati e la misura cross-correlation hanno un impatto significativo sulla sequenza cross-correlation prodotta.

Shape-based distance (SBD): Partendo dai ragionamenti precedenti viene scelto un coefficiente di standardizzazione che produca valori all'interno del range $[-1,1]$. Individuata come descritto la posizione ω dove $NCC_c(\vec{x}, \vec{y})$è massimizzata è possibile derivare la seguente misura di distanza tra $\vec{x}$ e $\vec{y}$:

$$SBD(\vec{x}, \vec{y}) = 1 - \max_{\omega}\left(\frac{CC_\omega(\vec{x}, \vec{y})}{\sqrt{R_0(\vec{x}, \vec{x}) \cdot R_0(1, \vec{y})}}\right)$$

Questa misurazione produrrà valori tra 0 e 2, dove con 0 avremo una perfetta similarità e con 2 una scarsa similarità. Fino ad ora abbiamo affrontato la shift-invariance. Per la scaling invariance viene trasformata ogni sequenza $\vec{x}$ in $\vec{x'} = \frac{\vec{x}-\mu}{\sigma}$ il che vuol dire che μ è zero e la sua deviazione standard σ è uno.

3.1.3 Calcolo dei centroidi

All'interno del refinement step, ad ogni iterazione, viene effettuato il calcolo dei centroidi dei cluster. Il centroide non è un oggetto del cluster ma può considerarsi piuttosto come il valore medio tra tutti gli elementi di un cluster. Molti task nell'analisi di time-series si basano sui metodi di sintesi di più time-series in una unica. Il metodo più semplice per estrarre il centroide di un cluster è calcolare, per ogni coordinata della serie temporale, un valore medio dato dalla media aritmetica dei valori corrispondenti di ogni sequenza. Questo approccio è usato ad esempio da K-Means. Si può dedurre quindi come l'estrazione dei centroidi dipenda criticamente dalla scelta della misura di distanza. In K-Shape i centroidi vengono determinati invece con una Shape Base Distance (SBD). Il calcolo del centroide corrisponde alla ricerca della minima somma del quadrato delle distanze tra tutte le time series all'interno di un cluster. Come definito dallo "Steiner's sequence problem" [36] dato un cluster $p_j \in P$, il corrispondente centroide $\vec{c_j}$ è definito dalla formula:

$$\vec{c_j} = \underset{\vec{\omega}}{argmin} \sum_{\vec{x_l} \in p_j} dist(\vec{\omega}, \vec{x_l})^2, \vec{\omega} \in \mathbb{R}$$

Visto che la cross-correlation definisce la similarità, piuttosto che una distanza, possiamo riscrivere il problema come la massimizzazione del quadrato della similarità fra tutte le sequenze:

$$\begin{aligned} \vec{\mu_k^*} &= \underset{\vec{\mu_k}}{argmax} \sum_{\vec{x_l} \in P_k} NCC_c(\vec{x_l}, \vec{\mu_k})^2 \\ &= \underset{\vec{\mu_k}}{argmax} \sum_{\vec{x_l} \in P_k} \left(\frac{max_\omega CC_\omega(x_l, \vec{\mu_k})}{\sqrt{R_0(\vec{x_l}, \vec{x_l}) \cdot R_0(\vec{\mu_k}, \vec{\mu_k})}}\right)^2 \end{aligned}$$

La misura richiede il calcolo dello shift ottimo per ogni $(\vec{x_i}) \in P_k$.

3.1.4 Complessità

La complessità dell'algoritmo K-Shape scala linearmente con il numero di time-series. Analizziamo la complessità ipotizzando di indicare con n il numero time-series, con k il numero di clusters e con m la lunghezza delle serie. Nell'assignment step k-shape calcola la dissimilarità di n serie verso k centroidi. Il calcolo della distanza richiede una complessità di $O(m \cdot \log(m))$ quindi in totale in questo step avremo una complessità di $O(n \cdot k \cdot m \cdot \log(m))$. Nel refinement step, per ogni cluster, viene calcolata la distanza tra le serie e i centroidi che ha una complessità di $O(m^3)$ quindi in totale in questo step avremo una complessità di $O(max\{n \cdot m^2, k \cdot m^3\})$. In totale quindi K-Shape ha, per ogni iterazione, una complessità di $O(max\{n \cdot k \cdot m \cdot \log(m), n \cdot m^2, k \cdot m^3\})$. Da qui vediamo la dipendenza lineare con il numero di serie e che la maggior parte del costo computazionale dipende dalla lunghezza delle serie. Nella maggior parte delle applicazioni però la lunghezza delle serie è molto minore del numero di serie, quindi non viene considerata un vincolo.

3.1.5 Risultati su dataset di esempio

Come anticipato gli autori del paper hanno condotto una valutazione sperimentale dell'algoritmo comparando i risultati con le misure di distanza e gli approcci di clustering più utilizzate allo stato dell'arte per le serie temporali. L'obiettivo è dimostrare l'efficacia della misura di distanza e del k-Shape. I risultati mostrano che la misura della distanza è migliore della distanza Euclidea e raggiunge risultati simili al costrained Dynamic Time Warping (cDTW - variante di DTW), considerata la migliore misura di distanza. A differenza della Shape Based Distance utilizzata con il k-Shape, cDTW richiede però un processo di tuning e impiega un ordine di grandezza di tempo in più. Nella tabella seguente si mostrano i risultati della ricerca sperimentale su 48 datasets di esempio[1]. Vengono paragonate le performance delle migliori misure di distanza allo stato dell'arte prendendo l'Euclidean Distance come riferimento. Nelle colonne "¿", "=" e "¡" troviamo

il numero di datasets in cui la misura di distanza in esame risulta migliore, uguale o peggiore rispetto ad ED. La colonna "Better" mostra se una distanza è migliore di ED da un punto di vista statistico. "Average accuracy" mostra l'accuratezza media raggiunta e "Runtime" msostra quanto la misura di distanza è più lenta rispetto a ED.

Distance Measure	>	=	<	Better	Average Accuracy	Runtime
DTW **DTW**$_{LB}$	29	2	17	✓	0.788	15573x 6040x
cDTW^{opt} **cDTW**$^{opt}_{LB}$	31	15	2	✓	0.814	2873x 322x
cDTW^{5} **cDTW**$^{5}_{LB}$	34	3	11	✓	0.809	1558x 122x
cDTW^{10} **cDTW**$^{10}_{LB}$	33	1	14	✓	0.804	2940x 364x
SBD$_{NoFFT}$ **SBD**$_{NoPow2}$ **SBD**	30	12	6	✓	0.795	224x 8.7x 4.4x

Figura 3.2. Comparazione delle misure di distanza[1]

La tabella mostra come SBD raggiunga ottimi risultati in termini di accuratezza mantenendo il costo computazionale più basso. Per il clustering viene mostrato che k-Shape è migliore di tutti gli algoritmi di tipo scalabili e non scalabile in termini di accuratezza, con la sola eccezione di k-medoids con cDTW. Questa misura di distanza però, come già detto, ha alcuni svantaggi che con k-Shape possono essere evitati. Nella seguente tabella vengono presentati i risultati della comparazione tra i migliori algoritmi di clustering scalabili, fra cui il k-Shape, con il k-AVG con Euclidean Distance. Nelle colonne "¿", "=" e "¡" troviamo il numero di datasets in cui l'algoritmo risulta migliore, uguale o peggiore rispetto a k-AVG+ED. La colonna "Better" e "Worse" mostrano se l'algoritmo è migliore o peggiore, dal punto di vista statistico, rispetto a k-AVG+ED. "Rand Index" mostra l'accuratezza raggiunta nei 48 datasets mentre "Runtime" mostra quanto l'algoritmo di clustering è più lento rispetto a k-AVG+ED. Dai risultati vediamo come solo k-Shape può considerarsi migliore di k-AVG+ED.

Algorithm	>	=	<	Better	Worse	Rand Index	Runtime
k**-AVG+SBD**	32	1	15	✗	✗	0.745	3.6x
k**-AVG+DTW**	10	0	38	✗	✓	0.584	3444x
KSC	22	0	26	✗	✗	0.636	448x
k**-DBA**	18	0	30	✗	✗	0.733	3892x
k**-Shape+DTW**	19	1	28	✗	✗	0.698	4175x
k**-Shape**	36	1	11	✓	✗	0.772	12.4x

Figura 3.3. Comparazione algoritmi scalabili[1]

Per avere un quadro completo nella prossima tabella vediamo i risultati della comparazione tra i migliori algoritmi di clustering non scalabili con il k-AVG con Euclidean Distance.

Al termine di queste analisi gli autori sottolineano come tra tutti gli algoritmi l'unico che raggiunge, ma non supera, risultati simili al k-Shape è PAM+cDTW che però presenta due svantaggi: 1) la sua misura di distanza richiede un processo d tuning per raggiungere buone performance; 2) il calcolo della matrice di dissimilarità richiede uno sforzo computazionale elevato. Per queste ragioni k-Shape viene considerato un algoritmo domain independent (queste performance infatti sono state testate su 48 datasets di diverso tipo), accurato e scalabile.

Algorithm	>	=	<	Better	Worse	Rand Index
H-S+ED	3	1	44	✗	✓	0.328
H-S+cDTW	7	0	41	✗	✓	0.371
H-S+SBD	6	1	41	✗	✓	0.349
H-A+ED	3	1	44	✗	✓	0.599
H-A+cDTW	9	0	39	✗	✓	0.617
H-A+SBD	8	0	40	✗	✓	0.541
H-C+ED	8	0	40	✗	✓	0.690
H-C+cDTW	15	0	33	✗	✓	0.699
H-C+SBD	17	0	31	✗	✓	0.697
S+ED	7	1	40	✗	✓	0.602
S+cDTW	18	1	29	✗	✓	0.563
S+SBD	38	0	10	✓	✗	0.769
PAM+ED	30	1	17	✗	✗	0.762
PAM+cDTW	38	1	9	✓	✗	0.772
PAM+SBD	35	1	12	✓	✗	0.780

Figura 3.4. Comparazione algoritmi non scalabili[1]

3.2 Pipeline di analisi

Vediamo ora come è stato utilizzato all'interno di questo lavoro l'algoritmo di clustering descritto. L'obiettivo che ci si pone in questo lavoro è duplice:

1. individuare i cluster all'interno del nostro dataset;
2. individuare i singoli titoli con andamento simile tra loro.

Inoltre ci si pone anche l'obiettivo di determinare dei parametri quantitativi che descrivano gli scenari analizzati. In particolare si vuole: (a) definire una misura di similarità intra-cluster al fine di quantificare quanto i titoli appartenenti ad un cluster siano simili tra loro; (b) definire una misura di similarità tra singoli titoli: al fine di quantificare quanto due titoli siano simili tra loro.

L'algoritmo creato si divide in 4 fasi:

1. Definizione parametri di input;
2. Import dei dati;
3. Clustering e definizione matrice di similarità;
4. Output:

   Definizione clusters;

   Similarità tra singoli oggetti (serie temporali);

   Drill-down su cluster o sub-set di oggetti(serie temporali).

Il seguente schema mostra le varie fasi dell'algoritmo:

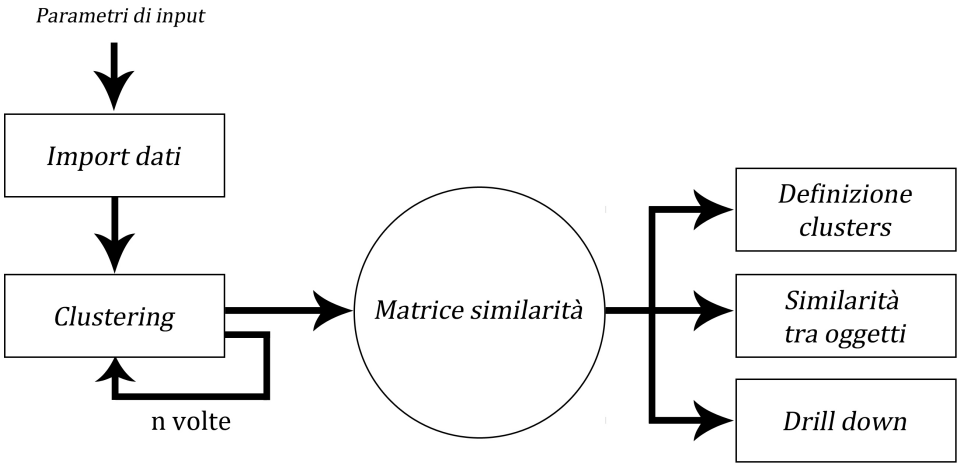

Figura 3.5. Descrizione pipeline di analisi

Vediamo ora la libreria utilizzata per il clustering e per la standardizzazione. Successivamente vedremo i dettagli delle singole fasi dell'algoritmo. Infine definiremo alcune misure quantitative che si possono derivare dai risultati e che possono essere utili a valutarli.

3.2.1 Libreria

L'algoritmo è stato scritto interamente in linguaggio Python 3.7.0 utilizzando per il clustering la libreria creata da Jörg Thalheim[37] che implementa fedelmente l'algoritmo descritto all'interno del paper precedentemente approfondito[1]. La funzione per il calcolo dei cluster è definita come segue:

$$kshape(x, k);$$

dove abbiamo in x le serie temporali e in k il numero di cluster da creare. La funzione restituisce una lista di coppie di valori in cui ogni coppia rappresenterà uno dei cluster individuati. Per ogni coppia avremo come primo valore il centroide del cluster e come secondo valore una lista con l'indice delle serie temporali appartenenti al cluster. L'algoritmo restituirà in ogni caso una lista con k elementi, ma alcuni di questi potrebbero essere vuoti (cluster vuoti). Nella funzione le serie temporali in input x dovranno essere, se necessario, già standardizzate. A questo scopo verrà utilizzata una funzione, appartenente alla libreria Scipy, definita come segue:

$$zscore(a, axis = 0, ddof = 0);$$

Questa funzione calcola la standardizzazione di ogni valore a in input. Trasforma cioè l'input da essere una variabile aleatoria (con media μ' e varianza σ^2) a diventare una variabile aleatoria con distribuzione standard, cioè con media zero e varianza uguale a 1. I parametri in input sono:

- a : dati di input;
- axis=0 : asse sul quale operare (default 0);
- ddof=0 : gradi di libertà sul calcolo della deviazione standard (default 0).

Vediamo un'applicazione d'esempio della funzione $kshape(x, k)$ con $zscore(a, axis = 0, ddof = 0)$:

```
from kshape.core import kshape, zscore

x = [[1,2,3,4], [0,1,2,3], [0,1,2,3], [1,2,2,3]]
k = 2
clusters = kshape(zscore(x, axis=1), k)

# RISULTATO:
#   [(array([-1.161895  , -0.38729833,  0.38729833,  1.161895  ]), [0, 1, 2]),
#    (array([-1.22474487,  0.        ,  0.        ,  1.22474487]), [3])]
```

Figura 3.6. Esempio applicazione libreria

3.2.2 Definizione parametri di input

Nella parte iniziale dell'algoritmo è necessario effettuare il setting dei parametri. Vediamo ogni parametro cosa ci permette di configurare:

- n: numero di volte che verrà eseguito l'algoritmo di clustering sugli stessi dati;
- k: numero di cluster da creare che verrà dato in input alla funzione $kshape(x, k)$;
- *date*: date (in termini di "anno" oppure di "anno-mese") oggetto di analisi;
- *somiglianza_threshold*: valore di somiglianza oltre la quale un oggetto verrà considerato simile ad un altro. Date n esecuzioni dell'algoritmo di clustering definiamo "somiglianza" il numero di volte in cui un oggetto verrà assegnato allo stesso cluster;
- *percentuale*: percentuale di oggetti alla quale un oggetto dovrà essere simile per considerarsi appartente al cluster;

- *numero_minimo_titoli*: numero minimo di titoli per formare un cluster;
- *path_dataset*: percorso in cui si trovano i file del dataset;
- *path_output*: percorso in cui depositare i file di output.

In questo momento alcuni parametri non saranno ancora chiari. Il loro utilizzo verrà esposto in modo completo nei paragrafi successivi.

3.2.3 Import dei dati

Il dataset è formato da file con la naming convention "XXXX_history_prices.csv" con al posto delle "XXXX" il simbolo del titolo. Questi file saranno all'interno del percorso definito nel parametro "*path_dataset*". L'algoritmo accede al percorso e apre tutti i file al suo interno. Per ogni file seleziona solo le righe dell'intervallo di tempo che si vuole esaminare che viene definito con il setting del parametro di input "*date*". Vista la grande differenza dei valori delle azioni per ogni titolo (ad esempio il titolo AMZN - Amazon.com ha una quotazione di circa 1700 $, mentre JPM - JPMorgan Chase & Co. ha una quotazione di circa 120 $) si è deciso di derivare dai dati grezzi un valore che fosse uniforme per tutti i titoli e che rendesse possibili le successive analisi. Si è scelto di calcolare la variazione percentuale giornaliera intesa come la variazione percentuale tra i prezzi di chiusura. Come prezzi di chiusura si è scelto di usare i valori di "*AdjClose*" in quanto, avendo subito un adeguamento che incorpora eventali variazioni di prezzo dovute a frazionamenti azionari o dividendi, risulta più efficace in analisi di questo tipo. Una eventuale variazione di prezzo dovuta ad un dividendo in un determinato giorno infatti aumenterebbe la distanza tra la serie e gli altri titoli pur non essendo indice, in linea generale, di un movimento contrastante con le altre serie, ma solo il risultato di una normale operazione finanziaria. Quindi per ogni riga di ogni file relativa all'intervallo di tempo selezionato verrà calcolata la variazione percentuale giornaliera rispetto al giorno precedente e il valore verrà inserito in una matrice in cui ogni riga descriverà l'andamento di un titolo. Questa matrice, una volta standardizzata con la funzione zscore, sarà l'input del nostro algoritmo di clustering.

3.2.4 Clustering e definizione matrice di similarità

Eseguito l'import dei dati vediamo ora come applicare l'algoritmo di clustering. Il K-Shape ha due svantaggi per la nostra applicazione. Per prima cosa non è deterministico: si basa infatti su un'assegnazione iniziale delle serie ai cluster completamente randomica. Partendo quindi dallo stesso input potrebbe darci in output dei cluster differenti. Inoltre l'algoritmo assegna tutte le serie ad un cluster quindi non contempla la possibilità che un oggetto possa non fare parte di nessun cluster. Per utilizzare l'algoritmo nel nostro campo applicativo si è scelto quindi di eseguirlo n volte, con n definito come parametro di input. In questo modo è diventato possibile:

- Definire i cluster con più robustezza considerando nella definizione dei cluster quante volte, nelle n esecuzioni, degli oggetti venivano assegnati allo stesso cluster;
- Definire una misura di distanza tra oggetti data dal numero di volte che un oggetto viene assegnato allo stesso cluster dell'altro;
- Escludere dall'analisi gli oggetti che, per quella fascia temporale, non possono considerarsi parte di un cluster;

Vedremo nel prossimo capitolo come sono stati affrontati questi punti nel dettaglio. Per gestire le n esecuzioni dell'algoritmo e per tenerne traccia è stata creata una matrice, che è stata chiamata "Matrice di similarità". La matrice di similarità ha i simboli che identificano gli oggetti (le singole serie temporali ricevute in input dall'algoritmo) in ascissa e in ordinata. Nelle celle all'interno della matrice avremo il numero di volte (compreso tra 0 e n) che, nelle n esecuzioni, i due oggetti sono stati assegnati allo stesso cluster. Questo parametro verrà chiamato "somiglianza". In Tabella 3.1 vediamo un esempio di una porzione di una matrice di similarità (con $n = 100$).

Osservando per esempio la prima riga, vediamo che nelle n esecuzioni il titolo ABMD è stato assegnato 74 volte allo stesso cluster di ABT, 16 volte allo stesso cluster di ACN, etc.

Symbol	ABMD	ABT	ACN	ADBE	ADI	ADM	ADP	ADSK	...
ABMD		74	16	20	5	0	1	4	...
ABT	74		4	6	3	0	0	3	...
ACN	16	4		77	15	0	31	14	...
ADBE	20	6	77		23	0	20	21	...
ADI	5	3	15	23		1	2	66	...
ADM	0	0	0	0	1		1	5	...
ADP	1	0	31	20	2	1		1	...
ADSK	5	11	2	4	9	9	0		...
...	...	...	...	...	...	...	...	...	

Tabella 3.1. Porzione di matrice di similarità con n=100

3.2.5 Definizione clusters

È stato quindi necessario definire un algoritmo che, data la matrice di similarità, estraesse i cluster. Nella matrice infatti troviamo solo il numero di volte che due titoli sono stati assegnati allo stesso cluster. A questo scopo utilizzeremo i parametri "*somiglianza_threshold*", "*percentuale*" e "*numero_minimo_titoli*" definiti inizialmente. L'obiettivo del calcolo dei cluster è estrarre dalla matrice di similarità i titoli che sono stati assegnati allo stesso cluster con maggiore frequenza. Supponiamo di avere in input 10 serie temporali, valore di $n = 100$, $somiglianza\_threshold = 85$, $percentuale = 70$ e $numero\_titoli = 3$ e ipotizziamo di avere la seguente matrice di similarità:

Symbol	A	B	C	D	E	F	G	H	I	J
A		0	10	1	4	90	7	40	33	1
B	0		5	6	7	86	12	0	4	1
C	10	5		99	1	11	8	98	85	88
D	1	6	99		3	2	1	99	85	1
E	4	7	1	3		8	97	4	8	5
F	90	86	11	2	8		5	35	25	0
G	7	12	8	1	97	5		22	33	8
H	40	0	98	99	4	35	22		86	3
I	33	4	85	85	8	25	33	86		2
J	1	1	88	1	5	0	8	3	2	

Tabella 3.2. Matrice di similarità di esempio con 10 serie e n=100

A questo punto l'algoritmo scorre tutta la matrice calcolando un possibile cluster per ogni riga. Per ogni riga si andranno ad estrarre tutti gli oggetti con somiglianza maggiore del parametro "*somiglianza_threshold*" e si metteranno in una lista insieme all'oggetto che identifica la riga. Ogni oggetto inserito in lista sarà rappresentato da una coppia di valori che sono il simbolo e il numero di occorrenze di questo simbolo in questo cluster. Per la prima riga avremo quindi:

$$list(A) = \{(A,1); (F,1)\}$$

Successivamente si andranno a scorrere tutti gli elementi di questa lista. Per ognuno di essi, ad eccezione dell'elemento già analizzato, si leggerà la corrispondente riga all'interno della matrice di similarità e si estrarranno solo gli oggetti con somiglianza maggiore del parametro "*somiglianza_threshold*". Gli oggetti estratti verranno aggiunti alla lista definita precedentemente oppure, se già presenti, ne verrà incremenetato il parametro indicante il numero di occorrenze. Verrà incrementato anche il parametro relativo al simbolo preso in analisi. Per il nostro caso si andrà quindi ad analizzare l'oggetto F e la lista $list(A)$ verrà aggiornata come segue:

$$list(A) = \{(A,2); (F,2); (B;1)\}$$

A questo punto si continuerà l'analisi scorrendo ancora la lista $list(A)$ e analizzando la riga relativa a B. La lista si aggiornerà come segue:

$$list(A) = (A,2); (F,3); (B;2)$$

Finiti gli elementi all'interno della lista da analizzare si procederà a verificare due condizioni:

1. Estrarre dalla lista solo gli elementi che hanno un numero di occorrenze rispetto al numero totale di titoli della lista maggiore, in termini percentuali, del valore "percentuale";
2. Verificare che il numero degli elementi estratti sia maggiore del "numero_minimo_titoli".

Nel nostro caso la prima condizione selezionerà solo i titoli che avranno un numero di occorrenze maggiore di 2. Infatti il numero dovrà essere maggiore del 70% (parametro "percentuale") del numero totale di titoli nella lista (3) che è pari a 2,1. La lista diventa quindi:

$$list(A) = (F,3)$$

La seconda condizione ci dice che il numero di titoli minimo per definire un cluster è 3. Nel nostro caso il cluster sarebbe formato da un solo elemento quindi possiamo dedurre che dall'oggetto A non è possibile estrarre un cluster:

$$list(A) = \{\emptyset\}$$

Effettuiamo rapidamente gli stessi calcoli per l'oggetto C. La lista definita inizialmente sarà:

$$list(C) = \{(C,1); (D,1); (H,1); (I,1)\}$$

Dopo aver analizzato la similarità di ogni titolo al suo interno la lista diventa:

$$list(C) = \{(C,5); (D,4); (H,4); (I,4); (J;1)\}$$

Dopo la verifica della prima e della seconda condizione la lista diventa:

$$list(C) = \{(C,5); (D,4); (H,4); (I,4)\}$$

Da cui deriviamo che il cluster formato dall'oggetto C è composto dagli elementi $\{C, D, H, I\}$. Dopo l'estrazione del cluster relativo ad ogni oggetto avremo sicuramente dei duplicati. Prendendo ad esempio il cluster derivato dall'analisi dell'oggetto C è facilmente deducibile che la stessa lista verrà estratta anche durante l'analisi degli altri elementi del cluster (D, H, I). Quindi infine verranno eliminati i cluster duplicati e verrà esportato il risultato all'interno del percorso definito dal parametro "*path_output*". Questo approccio, come abbiamo visto anche dall'esempio, farà si che non tutti i titoli verranno assegnati ad un cluster. L'algoritmo infatti restituisce solo i cluster che soddisfino i parametri stabiliti inizialmente.

3.2.6 Similarità intra-cluster

Definiamo ora come calcolare per ogni cluster un parametro che indichi il valore di similarità interna. Per fare questo prendiamo come esempio il cluster $\{C, D, H, I\}$ e consideriamo solo le righe e le colonne della matrice di somiglianza dei titoli oggetto di analisi. Nel nostro esempio la matrice diventa quella visibile in Tabella 3.3.

Symbol	**C**	**D**	**H**	**I**
C		99	98	85
D	99		99	85
H	98	99		86
I	85	85	86	

Tabella 3.3. Porzione di matrice di similarità per il calcolo della similarità intra-cluster

La similarità intra-cluster è stata definita calcolando la media dalla somiglianza di ogni titolo verso tutti gli altri titoli all'interno del cluster e poi calcolando la media di questi valori. Nell'esempio la media della somigliaza per ogni titolo è descritta in Tabella 3.4.

La media di questi valori sarà pari a 92,00 . Questo valore è il valore di similarità intra-cluster.

Symbol	C	D	H	I
Media somiglianza	94,00	94,33	94,33	85,33

Tabella 3.4. Media somiglianza serie con le altre all'interno dello stesso cluster

3.2.7 Similarità tra oggetti

Oltre ai cluster è possibile estrarre dalla matrice di similarità anche un parametro che può essere usato per definire la distanza tra due singoli oggetti. I valori di somiglianza tra i titoli infatti rispondono alla domanda: "*Dopo aver eseguito n volte l'algoritmo di clustering quante volte due oggetti sono stati assegnati allo stesso cluster?*". Questo valore sarà vicino a n tanto più gli oggetti saranno simili, al contrario sarà più vicino a 0 se le serie non saranno simili. Riprendendo l'esempio precedente e analizzando la similarità dell'oggetto A rispetto agli altri oggetti possiamo ordinare la colonna relativa ad A in ordine decrescente in modo da evidenziare quali titoli sono più simili ad A e quali lo sono meno. Il risultato è visibile in Tabella 3.5.

Symbol	A
F	90
H	40
I	33
C	10
G	7
E	4
D	1
J	1
B	0

Tabella 3.5. Somiglianza rispetto ad una singola serie

3.2.8 Drill down su cluster o sub-set di oggetti

È possibile effettutare un altro tipo di analisi che, partendo da un generico gruppo di oggetti, vada ad analizzare i legami che ci sono all'interno. Questo genere di analisi potrebbe essere fatta sia con gli oggetti di un cluster sia considerando anche un generico gruppo di titoli. Partiamo dal presupposto che ci aspettiamo che le aziende che sono coinvolte negli stessi settori di mercato è più probabile che abbiano un comportamento simile tra loro. Data questa premessa potremmo ad esempio analizzare i titoli delle aziende di un determinato Settore finanziario per vedere quali siano le dinamiche all'interno dello stesso settore. Questo genere di analisi viene fatta al fine di identificare eventuali sub-cluster e la chiameremo "Drill-down".

A prescindere da come vengono selezionati i titoli per questo genere di analisi è stato definita una pipiline di estrazione degli eventuali sub-cluster. L'algoritmo parte dalla matrice di similarità e ne estrae solo le sole righe e colonne dei titoli oggetto di analisi, come già fatto in precedenza. Successivamente si esegue sulla matrice di similarità estratta l'algoritmo di identificazione dei cluster precedentemente descritto aumentando gradualmente i vincoli di similarità. Questo approccio permette di partire con cluster più ampi con una similarità intra-cluster più bassa fino ad arrivare a cluster meno numerosi ma con alta similarità interna. I titoli che non vengono assegnati a nessun cluster vengono considerati con comportamento non assimilabile a nessun altro titolo fra quelli in analisi.

Partendo da questa analisi è stato ideato anche un metodo di visualization che verrà illustrato all'interno del prossimo capitolo. Spesso infatti risulta difficoltoso presentare in modo esaustivo i risultati degli algoritmi di data mining. Inoltre è importante scegliere o ideare i metodi di visualization in base sia all'applicazione specifica sia a coloro da cui dovranno essere letti.

Capitolo 4

Esperimenti

All'interno di questa sezione andremo a vedere i risultati della metodologia descritta fino ad ora applicata al nostro dataset. Per ogni esperimento verrà inserita una tabella con i parametri di input utilizzati (ad eccezione dei parametri "*path_dataset*" e "*path_output*") come la Tabella 4.1.

Parametri di input:	
n	...
k	...
date	...
somiglianza_threshold	...
percentuale	...
numero_minimo_titoli	

Tabella 4.1. Parametri di input d'esempio

Inizieremo con una premessa riguardante il *Global Industry Classification Standard (GISC)* che sarà elemento chiave per la nostra analisi. Questo standard classifica i vari titoli in Sectors, Industry Groups, Industries o Sub-Industries. Vedremo che i cluster spesso saranno simili alla classificazione definita dallo standard. Utilizzeremo questo elemento non come ground truth ma piuttosto come mezzo di comprensione dei risultati ottenuti. Continueremo poi con le sperimentazioni vere e proprie iniziando con le modalità di definizione del parametro k (numero di cluster), del parametro *date* (archi temporali di analisi) e dei parametri *somiglianza_threshold*, *percentuale* e *numero_minimo_titoli* (somiglianza cluster). Vedremo poi i risultati dell'applicazione dell'algoritmo sul dataset, il calcolo della similarità intra-cluster, la similarità tra singoli oggetti. Infine verrà esplorata la tecnica di drill down sia su interi settori sia su cluster presentando anche un nuovo metodo di visualization dei dati.

4.1 Experimental design

Come già anticipato la metodologia e gli algoritmi descritti non hanno un costo computazionale molto alto grazie al dataset utilizzato (time-series non eccessivamente lunghe) e all'efficienza dell'algoritmo di clustering.

Per questo motivo per questi esperimenti si è scelto di utilizzare una macchina con le caratteristiche descritte in Figura 4.1.

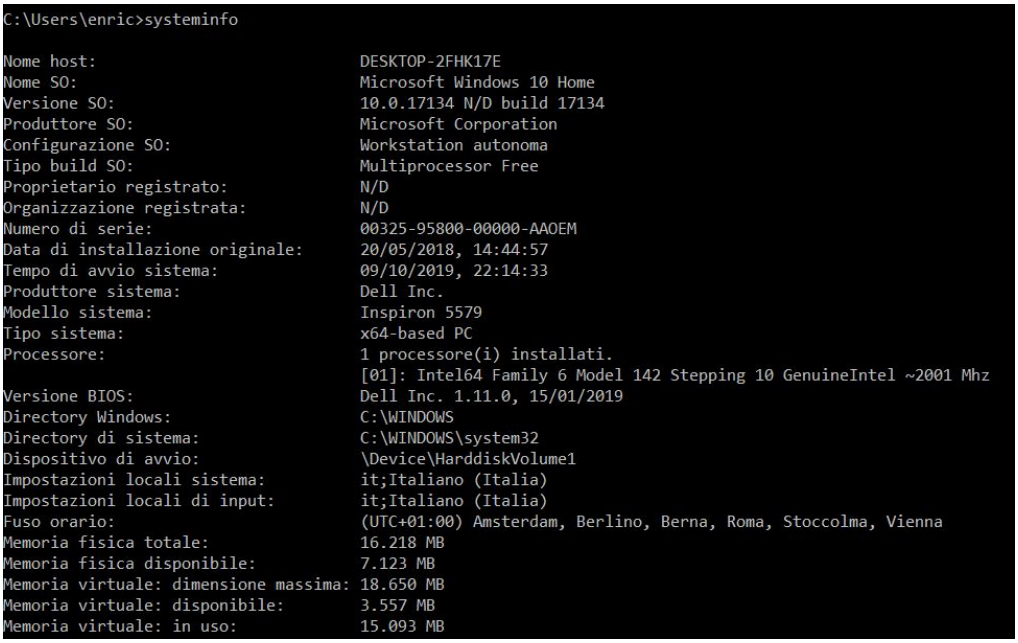

Figura 4.1. Informazioni di sistema macchina

Per valutare i tempi di esecuzione prendiamo come esempio una esecuzione sperimentale su un dataset formato da 500 serie temporali lunghe 252 campioni ciascuna con numero di cluster (parametro k) pari a 15 e 100 esecuzioni dell'algoritmo (parametro n). In questa situazione il tempo di esecuzione totale è stato di di $1217940ms$ (poco più di 20 minuti) di cui: $14740ms$ (circa 15 secondi) impiegati per l'import dei dati, $1201770ms$ (circa 20 minuti) per le 100 esecuzioni dell'algoritmo di clustering e $1430ms$ (circa 1,4 secondi) per l'estrazione dei cluster dalla matrice di similarità.

4.2 Dataset

Come già anticipato questo lavoro si basa sull'analisi di un dataset formato dai dati di andamento dei titoli azionari dell'indice S&P500. All'interno di questo indice, fondato da Standard & Poor's nel 1957, troviamo le 500 aziende statunitensi a maggiore capitalizzazione contrattate al New York Stock Exchange (Nyse), all'American Stock Exchange (Amex) e al Nasdaq. Viene ritenuto il più importante indice azionario nordamericano. Entrando nel dettaglio il dataset è composto da un file per ogni titolo azionario il cui nome segue la naming convention "XXXX_history_prices.csv" con al posto delle "XXXX" il simbolo del titolo. All'interno di ogni file si trova una riga per ogni giorno di contrattazione dal 03/01/2007 al 10/01/2018 quindi equivalente ad oltre 10 anni di contrattazioni su quel titolo. Ogni file avrà quindi un massimo di 2777 righe che equivalgono al numero di giorni di contrattazione nell'intervallo di date indicato. Alcuni titoli potranno avere meno righe in relazione ad un eventuale inizio delle contrattazioni successivo al 03/01/2007 o termine delle contrattazioni precedente al 10/01/2018. Il dataset è stato scaricato interamente tramite le API fornite dal portale Yahoo Finance[38]. Per ogni riga abbiamo i seguenti campi:

- **Open**: è il primo prezzo negoziato da un titolo all'apertura di un giorno di negoziazione;
- **High**: è il prezzo massimo raggiunto dal titolo durante le negoziazioni di una intera giornata;
- **Low**: è il prezzo minimo raggiunto dal titolo durante le negoziazioni di una intera giornata;
- **Close**: è l'ultimo prezzo negoziato da un titolo alla chiusura in un giorno di negoziazione;
- **Adj Close**: è l'ultimo prezzo negoziato da un titolo alla chiusura in un giorno di negoziazione dopo aver applicato degli adeguamenti che incorporino eventuali cambi di prezzo causati da frazionamenti azionari o dividendi. Il dato viene adeguato considerando appropriati moltiplicatori in accordo con gli standard definiti dal "Center for Research in Security Prices (CRSP)"[39];
- **Volume**: è il numero di azioni di un titolo negoziate in un giorno di negoziazione.

Vediamo, per esempio, un estratto del file "MSFT_history_prices.csv" relativo alle contrattazioni del titolo" Microsoft corporation" (MSFT) in Tabella 4.2.

Date,	**Open,**	**High,**	**Low,**	**Close,**	**Adj Close,**	**Volume**
2007-01-03,	29.910000,	30.250000,	29.400000,	29.860001,	22.574831,	76935100
2007-01-04,	29.700001,	29.969999,	29.440001,	29.809999,	22.537027,	45774500
2007-01-05,	29.629999,	29.750000,	29.450001,	29.639999,	22.408501,	44607200
... ,	... ,	... ,	... ,	... ,	... ,	... ,

Tabella 4.2. Estratto file su andamento titolo MSFT

Dal dataset viene estratta, come descritto precedentemente, la variazione percentuale giornaliera di ogni titolo. Vediamo un esempio relativo ai titoli A, AAL, AAP, AAPL nel il periodo compreso tra il 3 gennaio 2007 e il 12 gennaio 2007. I valori di "Adj Close" in questo range temporale sono descritti in Tabella 4.3

TITOLO	**03-01-07**	**04-01-07**	**05-01-07**	**08-01-07**	**09-01-07**	**10-01-07**	**11-01-07**	**12-01-07**
A	22,643923	22,716543	22,505285	22,426065	22,452469	22,247818	22,247818	22,181801
AAL	53,934589	56,367863	55,840988	55,496105	55,467361	56,45409	58,628719	58,255108
AAP	34,121075	34,341648	33,584034	33,699104	33,986813	34,03476	34,907444	34,715649
AAPL	8,01682	8,194759	8,136404	8,176582	8,855811	9,279611	9,164813	9,051928

Tabella 4.3. Estratto andamento di AdjClose sui titoli A, AAL, AAP, AAPL

Da cui si deriva la matrice con la variazione percentuale giornaliera in Tabella 4.4

Possiamo osservare in Figura 4.2 l'andamento di questo dato .

TITOLO	04-01-07	05-01-07	08-01-07	09-01-07	10-01-07	11-01-07	12-01-07
A	0,32070	-0,92997	-0,35200	0,1177	-0,91148	0	-0,29673
AAL	4,51152	-0,93470	-0,61761	-0,05179	1,77893	3,85203	-0,63724
AAP	0,64644	-2,20610	0,342633	0,85375	0,14107	2,56409	-0,54943
AAPL	2,21957	-0,71210	0,49380	8,30700	4,78555	-1,23709	-1,23172

Tabella 4.4. Estratto variazione percentuale di AdjClose sui titoli A, AAL, AAP, AAPL

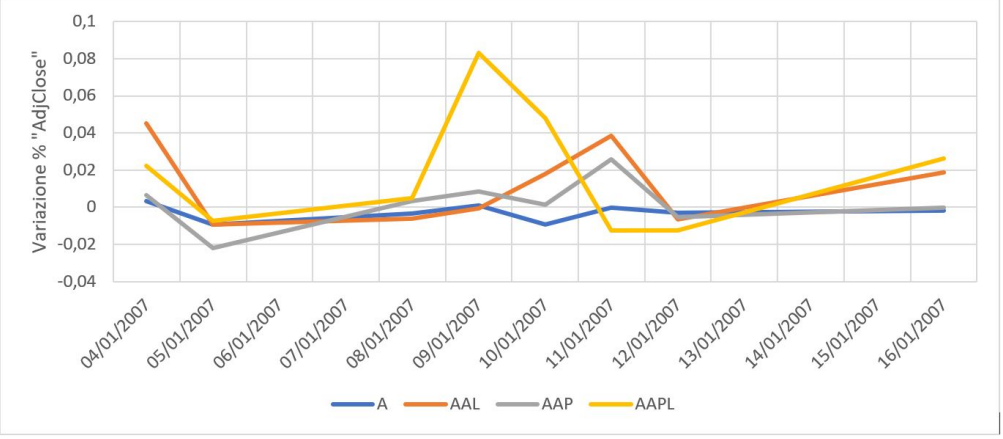

Figura 4.2. Andamento d'esempio variazione percentuale giornaliera

4.3 Global Industry Classification Standard

Nella lettura dei risultati dell'algoritmo, all'interno di questa sezione, si farà spesso riferimento ai settori industriali di appartenenza dei vari titoli azionari. I riferimenti seguono lo standard *GISC (Global Industry Classification Standard)*[40] che è un tipo di tassonomia industriale sviluppata nel 1999 da MSCI e Standard & Poor's che divide le aziende in gruppi industriali in base a simili processi produttivi e prodotti. La struttura GISC, con ultimo aggiornamento a Settembre 2018, è formata da 11 Sectors, 24 Industry Groups, 69 Industries e 158 Sub-Industries. La struttura viene definita sia da una scala gerarchica di numeri sia da una descrizione. Osservando ad esempio il settore "Energy" (codice 10) vediamo che è formato da un unico Industry Group (codice 1010), da due Industry (codici 101010 e 101020) e da 7 sub-industry.

Sector		Industry Group		Industry		Sub-Industry	
10	Energy	1010	Energy	101010	Energy Equipment & Services	10101010	Oil & Gas Drilling
						10101020	Oil & Gas Equipment & Services
				101020	Oil, Gas & Consumable Fuels	10102010	Integrated Oil & Gas
						10102020	Oil & Gas Exploration & Production
						10102030	Oil & Gas Refining & Marketing
						10102040	Oil & Gas Storage & Transportation
						10102050	Coal & Consumable Fuels

Figura 4.3. Esempio di settore GISC

Tra gli allegati di questo lavoro è possibile consultare l'ultima pubblicazione della struttura GISC completa[40].

4.4 Definizione parametro k

Come visto precedentemente la tecnica di clustering utilizzata ha bisogno in input del numero di cluster da formare. Per come è stato strutturato l'algoritmo questo valore definirà qual è il numero di cluster da creare in ognuno degli n cicli. Il numero di cluster finali sarà invece strettamente legato ai parametri "*somiglianza_threshold*", "*numero_minimo_titoli*" e "*percentuale*" e dalla matrice di somiglianza che verrà creata secondo la tecnica descritta in precedenza. Vedremo ora i risultati dell'applicazione sperimentale dell'algoritmo con diversi valori di k. Il range dei valori di k sarà $10 \leq k \leq 20$. Questi valori sono stati scelti partendo dalla considerazione che probabilmente ci saranno un numero di cluster simile al numero dei sector definiti dal GISC, cioè 11. La scelta degli altri parametri di input verrà discussa successivamente. Nella Tabella 4.5 vengono riassunti tutti i parametri di input utilizzati.

Parametri di input:	
n	100
k	$10 \leq k \leq 20$
date	2015
somiglianza_threshold	85
percentuale	70
numero_minimo_titoli	5

Tabella 4.5. Parametri di input per definizione parametro k

I cluster formati sono descritti nelle tabelle alle pagine successive. Riassumiamo in Tabella 4.6 i risultati in termini di "Media membri cluster", "Media somiglianza" e "Numero cluster" ottenuti al variare di k.

K	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
Media membri clusters	13,8	21,1	22,6	20,0	17,0	18,2	16,3	17,0	13,9	13,3	17,3
Media somiglianza	95,4	96,5	95,9	96,0	94,9	95,5	95,5	95,5	94,5	94,9	95,1
Numero clusters	13	9	8	9	11	10	11	10	12	10	8

Tabella 4.6. Media membri, somiglianza e numero cluster al variare di k

Come si può notare in Tabella 4.6 e nelle Figure 4.4 e 4.5 la variazione di k con i parametri di input scelti non determina drastici cambiamenti. Per quanto riguarda la media della somiglianza intra-cluster il parametro ha una variabilità molto bassa, questo è da considerarsi come diretta conseguenza del setting degli altri parametri di input. Per analizzare gli altri risultati si è scelto di premiare i valori di k che hanno generato più cluster e più grandi. Per questo motivo sono stati segnati in rosso nella tabella 4.6 i valori più bassi della media dei membri dei clusters e del numero dei clusters formati. Questo perché un algoritmo che genera pochi cluster probabilmente sta escludendo di poco alcuni cluster che invece potrebbero avere la loro rilevanza. Una media bassa dei membri dei cluster invece rileva l'esclusione dai cluster di alcuni oggetti, che invece con altri valori di k ne facevano parte pur non pregiudicando i valori di somiglianza intra-cluster. Nel caso specifico si può fare l'esempio di cluster non individuati, come quello formato da aziende di "Information Technology" o "Airlines", oppure di aziende escluse da un cluster, come per il cluster formato da aziende che si occupano di "Consumer Staples" che passa da 19 membri per $k = 15$ a 7 membri per $k = 18$. Naturalmente i risultati ottenuti sono influenzati anche da tutti gli altri parametri. Per questo motivo per le analisi future si è scelto di mantenere il valore di k all'interno del range $14 \leq k \leq 17$ verificando i risultati ottenuti al variare degli altri parametri.

K=10			K=11			K=12			K=13			K=14			K=15		
Clusters	N°	S.	Clusters	N°	S.	Clusters	N°	S.	Clusters	N°	S.	Clusters	N°	S.	Clusters	N°	S.
AAL, ALK, DAL, LUV, UAL	5	98,6							AAL, ALK, DAL, LUV, UAL	5	95,2	AAL, ALK, DAL, LUV, UAL	5	94,6	AAL, ALK, DAL, LUV, UAL	5	92,3
AAP, AZO, BBY, DG, DLTR, FL, GPC, GPS, JWN, KSS, LB, M, ORLY, PVH, RL, ROST, TGT, TJX, TPR, TSCO, ULTA, VFC	22	93,8	BBY, DG, DLTR, FL, GPS, HBI, JWN, KSS, LB, LOW, M, NKE, PVH, ROST, TGT, TJX, TPR, UAA, ULTA, VFC	20	94,6	BBY, DG, DLTR, FL, GPS, HBI, JWN, KSS, LB, LOW, M, NKE, PVH, ROST, TGT, TJX, TPR, UAA, ULTA, VFC	20	94,1	BBY, DG, DLTR, FL, GPS, JWN, KSS, LB, M, NKE, PVH, ROST, TGT, TJX, TPR, UAA, ULTA, VFC	18	93,7	BBY, DG, DLTR, FL, GPS, JWN, KSS, LB, M, NKE, ROST, TGT, TJX, UAA, ULTA, VFC	16	94,5	BBY, DG, DLTR, FL, GPS, JWN, KSS, LB, M, ROST, TGT, TJX, ULTA, VFC	14	96,2
AEE, AEP, AES, AWK, CMS, CNP, D, DTE, DUK, ED, EIX, ES, ETR, EXC, FE, LNT, NEE, NI, PCG, PEG, PNW, PPL, SCG, SO, SRE, WEC, XEL	27	98,8	AEE, AEP, AWK, CMS, CNP, D, DTE, DUK, ED, EIX, ES, ETR, EXC, FE, LNT, NEE, NI, PCG, PEG, PNW, PPL, SCG, SO, SRE, WEC, XEL	26	100	AEE, AEP, AWK, CMS, CNP, D, DTE, DUK, ED, EIX, ES, ETR, EXC, FE, LNT, NEE, NI, PCG, PEG, PNW, PPL, SCG, SO, SRE, WEC, XEL	26	99,8	AEE, AEP, AWK, CMS, CNP, D, DTE, DUK, ED, EIX, ES, ETR, EXC, FE, LNT, NEE, NI, PCG, PEG, PNW, PPL, SCG, SO, SRE, WEC, XEL	26	99,5	AEE, AEP, AWK, CMS, CNP, D, DTE, DUK, ED, EIX, ES, ETR, EXC, FE, LNT, NEE, NI, PCG, PEG, PNW, PPL, SCG, SO, SRE, WEC, XEL	26	99,2	AEE, AEP, AWK, CMS, CNP, D, DTE, DUK, ED, EIX, ES, ETR, EXC, FE, LNT, NEE, NI, PCG, PEG, PNW, PPL, SCG, SO, SRE, WEC, XEL	26	99,9
AFL, AJG, ALL, CB, CINF, HIG, L, PGR, RE, TRV	10	93,2	AON, BRK-B, CB, CINF, MMC, PGR, RE, TRV	8	94,4				AJG, AON, BRK-B, CB, CINF, MMC, PGR, RE, TRV	9	92,5	AON, BRK-B, CB, CINF, MMC, PGR, RE, TRV	8	91,6	AJG, AON, BRK-B, CB, CINF, MMC, PGR, RE, TRV	9	90,9
AIV, ARE, AVB, DRE, EQR, ESS, EXR, HCP, IRM, MAA, O, PLD, PSA, UDR, VTR, WELL	16	98,6	AIV, ARE, AVB, BXP, DLR, DRE, EQR, ESS, EXR, FRT, HCP, IRM, KIM, MAA, O, PLD, PSA, REG, SLG, SPG, UDR, VNO, VTR, WELL	24	99,1	AIV, ARE, AVB, BXP, DLR, DRE, EQR, ESS, EXR, FRT, HCP, IRM, KIM, MAA, O, PLD, PSA, REG, SLG, SPG, UDR, VNO, VTR, WELL	24	99,5	AIV, ARE, AVB, BXP, DLR, DRE, EQR, ESS, EXR, FRT, HCP, IRM, KIM, MAA, O, PLD, PSA, REG, SLG, SPG, UDR, VNO, VTR, WELL	24	99,5	AIV, ARE, AVB, BXP, DLR, DRE, EQR, ESS, EXR, FRT, HCP, IRM, KIM, MAA, O, PLD, PSA, REG, SLG, SPG, UDR, VNO, VTR, WELL	24	99,3	AIV, ARE, AVB, BXP, DLR, DRE, EQR, ESS, EXR, FRT, HCP, IRM, KIM, MAA, O, PLD, PSA, REG, SLG, SPG, UDR, VNO, VTR, WELL	24	99,3
KIM, MAC, SLG, SPG, VNO	5	97,6															
AMG, BEN, BLK, IVZ, TROW	5	91,5															
APA, APC, BHGE, COG, COP, CVX, CXO, DVN, EOG, FANG, FCX, FTI, HAL, HES, HP, KMI, LYB, MRO, NBL, NEM, NFX, NOV, NRG, OKE, OXY, PXD, SLB, WMB, XEC, XOM	26	97,1	APA, APC, BHGE, COG, COP, CVX, CXO, DVN, EOG, FANG, FCX, FTI, HAL, HES, HP, KMI, LYB, MRO, NBL, NEM, NFX, NOV, NRG, OKE, OXY, PXD, SLB, WMB, XEC, XOM	30	98,7	APA, APC, BHGE, COG, COP, CVX, CXO, DVN, EOG, FANG, FCX, FTI, HAL, HES, HP, KMI, LYB, MRO, NBL, NEM, NFX, NOV, NRG, OKE, OXY, PXD, SLB, WMB, XEC, XOM	30	97,1	APA, APC, BHGE, COG, COP, CVX, CXO, DVN, EOG, FANG, FCX, FTI, HAL, HES, HP, KMI, LYB, MRO, NBL, NEM, NFX, NOV, NRG, OKE, OXY, PXD, SLB, WMB, XEC, XOM	30	97,4	APA, APC, BHGE, COG, COP, CVX, CXO, DVN, EOG, FANG, FTI, HAL, HES, HP, KMI, MRO, NBL, NEM, NFX, NOV, OKE, OXY, PXD, SLB, WMB, XEC, XOM	27	97,7	APA, APC, BHGE, COG, COP, CVX, CXO, DVN, EOG, FANG, FCX, FTI, HAL, HES, HP, KMI, LYB, MRO, NBL, NFX, NOV, NRG, OKE, OXY, PXD, SLB, WMB, XEC, XOM	29	96,6
BAC, BBT, BK, BRK-B, C, CFG, CMA, CME, COF, DFS, ETFC, FITB, GS, HBAN, JPM, KEY, LNC, MET, MS, MTB, NTRS, PBCT, PFG, PNC, PRU, RF, RJF, SCHW, SIVB, STI, STT, TMK, UNM, USB, WFC, ZION	36	95,11	AIG, AIZ, AMP, BAC, BBT, BK, C, CFG, CMA, CME, COF, ETFC, FITB, GS, HBAN, HIG, JPM, KEY, LNC, MET, MS, MTB, NTRS, PBCT, PFG, PNC, PRU, RF, RJF, SCHW, SIVB, STI, STT, TMK, UNM, USB, WFC, ZION	38	97,18	AIG, AIZ, AMP, BAC, BBT, BK, C, CFG, CMA, CME, COF, ETFC, FITB, GS, HBAN, HIG, JPM, KEY, LNC, MET, MS, MTB, NTRS, PBCT, PNC, PRU, RF, RJF, SCHW, SIVB, STI, STT, TMK, UNM, USB, WFC, ZION	38	95,5	AIG, AMP, BAC, BBT, BK, C, CFG, CMA, CME, COF, ETFC, FITB, GS, HBAN, JPM, KEY, LNC, MET, MS, MTB, NTRS, PBCT, PNC, PRU, RF, RJF, SCHW, SIVB, STI, STT, UNM, USB, WFC, ZION	34	98,03	AIG, AMP, BAC, BBT, BK, C, CFG, CMA, CME, COF, ETFC, FITB, GS, HBAN, JPM, KEY, LNC, MET, MS, MTB, NTRS, PBCT, PFG, PNC, PRU, RF, RJF, SCHW, SIVB, STI, STT, TMK, UNM, USB, WFC, ZION	36	96,3	BAC, BBT, BK, C, CFG, CMA, COF, ETFC, FITB, GS, HBAN, JPM, KEY, LNC, MET, MS, MTB, NTRS, PBCT, PNC, PRU, RF, RJF, SCHW, SIVB, STI, STT, UNM, USB, WFC, ZION	31	98,1
CAG, CPB, GIS, HRL, HSY, K, SJM	7	98,6	BF-B, CAG, CHD, CL, CLX, CPB, GIS, HRL, HSY, K, KMB, KO, MDLZ, MKC, MO, PEP, PG, PM, SJM, TAP	20	96,5	BF-B, CAG, CHD, CL, CLX, CPB, GIS, HRL, HSY, K, KMB, KO, MDLZ, MKC, MO, PEP, PG, PM, SJM, TAP	20	95,3	BF-B, CAG, CHD, CL, CLX, CPB, GIS, HRL, HSY, K, KMB, KO, MDLZ, MKC, MO, PEP, PG, PM, SJM, TAP	20	95,1	BF-B, CAG, CHD, CL, CLX, CPB, GIS, HRL, HSY, K, KMB, KO, MDLZ, MKC, MO, PEP, PG, PM, SJM, TAP	20	93,1	BF-B, CAG, CHD, CL, CLX, CPB, GIS, HRL, HSY, K, KMB, KO, MDLZ, MKC, MO, PEP, PG, PM, SJM	19	95,5
CBS, DIS, DISCA, DISCK, DISH, FOX, FOXA, T, VIAB	9	94,3															
CHRW, FDX, JBHT, NSC, UNP, UPS	6	89,8															
GD, LLL, LMT, NOC, RTN	5	93,8															
			ABC, AGN, ALGN, BAX, BDX, BMY, CAH, EW, LLY, MYL, PFE, PRGO, RMD, ZBH	14	92,1												
			AME, CMI, DE, EMR, ETN, GWW, PH, PNR, ROK, XYL	10	95,7	AME, CMI, DE, EMR, ETN, FAST, GWW, IR, PCAR, PH, PNR, ROK, XYL	13	92,3						CMI, DE, EMR, ETN, GWW, PH, PNR, ROK, XYL	9	92,7	CAT, CMI, DE, EMR, ETN, FAST, GWW, PH, PNR, ROK, XYL	11	94,5
						ADI, AVGO, INTC, KLAC, LRCX, MCHP, MXIM, NVDA, TXN, XLNX	10	93,8	ADI, AMAT, AVGO, INTC, KLAC, LRCX, MCHP, MU, MXIM, NVDA, QRVO, SWKS, TXN, XLNX	14	93,0	ADI, AMAT, INTC, KLAC, LRCX, MCHP, MU, MXIM, NVDA, TXN, XLNX	11	92,5	ADI, AVGO, FFIV, INTC, KLAC, LRCX, MCHP, MU, MXIM, NVDA, QRVO, SWKS, TXN, XLNX	14	91,2
												ARNC, DOV, FLR, FLS, PWR	5	91,9			

Figura 4.4. Clusters formati per k con valori $10 \leq k \leq 15$

K=16			K=17			K=18			K=19			K=20		
Clusters	N°	S.	Clusters	N°	S.	Clusters	N°	S.	Clusters	N°	S.	Clusters	N°	S.
AAL, ALK, DAL, LUV, UAL	5	98,4	AAL, ALK, DAL, LUV, UAL	5	97,3	AAL, ALK, DAL, LUV, UAL	5	97,2	AAL, ALK, DAL, LUV, UAL	5	96,8	AAL, ALK, DAL, LUV, UAL	5	97,2
BBY, DG, DLTR, FL, GPS, JWN, KSS, LB, M, ROST, TGT, TJX, ULTA, VFC	14	95	DG, DLTR, FL, GPS, JWN, KSS, LB, M, ROST, TGT, TJX, ULTA, VFC	13	96,3	DG, DLTR, FL, GPS, JWN, KSS, LB, M, ROST, TGT, TJX, ULTA	12	94,3	DG, DLTR, FL, GPS, JWN, KSS, LB, M, ROST, TGT, TJX, ULTA	12	92,4	DG, DLTR, FL, GPS, JWN, KSS, LB, M, ROST, TGT, TJX, ULTA, VFC	13	93,3
AEE, AEP, AWK, CMS, CNP, D, DTE, DUK, ED, EIX, ES, ETR, EXC, FE, LNT, NEE, NI, PCG, PEG, PNW, PPL, SCG, SO, SRE, WEC, XEL	26	99,5	AEE, AEP, AWK, CMS, CNP, D, DTE, DUK, ED, EIX, ES, ETR, EXC, FE, LNT, NEE, NI, PCG, PEG, PNW, PPL, SCG, SO, SRE, WEC, XEL	26	99,7	AEE, AEP, AWK, CMS, CNP, D, DTE, DUK, ED, EIX, ES, ETR, EXC, FE, LNT, NEE, NI, PCG, PEG, PNW, PPL, SCG, SO, SRE, WEC, XEL	26	98,7	AEE, AEP, AWK, CMS, CNP, D, DTE, DUK, ED, EIX, ES, ETR, EXC, FE, LNT, NEE, NI, PCG, PEG, PNW, PPL, SCG, SO, SRE, WEC, XEL	26	98,3	AEE, AEP, AWK, CMS, CNP, D, DTE, DUK, ED, EIX, ES, ETR, EXC, FE, LNT, NEE, NI, PCG, PEG, PNW, PPL, SCG, SO, SRE, WEC, XEL	26	98,6
BRK-B, CB, CINF, MMC, PGR, RE, TRV	7	91,5							AJG, CB, CINF, MMC, RE, TRV	6	90,7	BRK-B, CB, CINF, PGR, RE, TRV	6	90
AIV, ARE, AVB, BXP, DLR, DRE, EQR, ESS, EXR, FRT, HCP, IRM, KIM, MAA, O, PLD, PSA, REG, SLG, SPG, UDR, VNO, VTR, WELL	24	98,7	AIV, ARE, AVB, BXP, DLR, DRE, EQR, ESS, EXR, FRT, HCP, IRM, KIM, MAA, O, PLD, PSA, REG, SLG, SPG, UDR, VNO, VTR, WELL	24	98,5	AIV, ARE, AVB, BXP, DLR, DRE, EQR, ESS, EXR, FRT, HCP, IRM, KIM, MAA, O, PLD, PSA, REG, SLG, SPG, UDR, VNO, VTR, WELL	26	97,8	AIV, ARE, AVB, BXP, DLR, DRE, EQR, ESS, EXR, FRT, HCP, IRM, KIM, MAA, O, PLD, PSA, REG, SLG, SPG, UDR, VNO, VTR, WELL	24	98,4	AIV, ARE, AVB, BXP, DLR, DRE, EQR, ESS, EXR, FRT, HCP, IRM, KIM, MAA, MAC, O, PLD, PSA, REG, SLG, SPG, UDR, VNO, VTR, WELL	25	96,6
APA, APC, BHGE, COG, COP, CVX, CXO, DVN, EOG, FANG, FTI, HAL, HES, HP, KMI, MRO, NBL, NFX, NOV, OKE, OXY, PXD, SLB, WMB, XEC, XOM	26	97,7	APA, APC, BHGE, COG, COP, CVX, CXO, DVN, EOG, FANG, FTI, HAL, HES, HP, KMI, MRO, NBL, NFX, NOV, OKE, OXY, PXD, SLB, WMB, XEC, XOM	26	98,8	APA, APC, BHGE, COG, COP, CVX, CXO, DVN, EOG, FANG, FTI, HAL, HES, HP, KMI, MRO, NBL, NFX, NOV, OKE, OXY, PXD, SLB, WMB, XEC, XOM	26	97,6	APA, APC, BHGE, COG, COP, CVX, CXO, DVN, EOG, FANG, FTI, HAL, HES, HP, KMI, MRO, NBL, NFX, NOV, OKE, OXY, PXD, SLB, WMB, XEC, XOM	26	98	APA, APC, BHGE, COG, COP, CVX, CXO, DVN, EOG, FANG, FTI, HAL, HES, HP, KMI, MRO, NBL, NFX, NOV, OKE, OXY, PXD, SLB, WMB, XEC, XOM	26	97
AMP, BAC, BBT, BK, C, CFG, CMA, COF, ETFC, FITB, GS, HBAN, JPM, KEY, LNC, MET, MS, MTB, NTRS, PBCT, PNC, PRU, RF, RJF, SCHW, SIVB, STI, STT, UNM, USB, WFC, ZION	32	97,3	BAC, BBT, BK, C, CFG, CMA, COF, ETFC, FITB, GS, HBAN, JPM, KEY, LNC, MET, MS, MTB, NTRS, PBCT, PNC, PRU, RF, RJF, SCHW, SIVB, STI, STT, UNM, USB, WFC, ZION	31	94,1	BAC, BBT, BK, C, CFG, CMA, COF, ETFC, FITB, GS, HBAN, JPM, KEY, LNC, MET, MS, MTB, NTRS, PBCT, PNC, PRU, RF, RJF, SCHW, SIVB, STI, STT, UNM, USB, WFC, ZION	31	93,4	BAC, BBT, BK, C, CFG, CMA, COF, ETFC, FITB, GS, HBAN, JPM, KEY, LNC, MET, MS, MTB, NTRS, PBCT, PNC, PRU, RF, RJF, SCHW, SIVB, STI, STT, UNM, USB, WFC, ZION	31	95,6	BAC, BBT, BK, C, CFG, CMA, COF, ETFC, FITB, GS, HBAN, JPM, KEY, LNC, MET, MS, MTB, NTRS, PBCT, PNC, PRU, RF, RJF, SCHW, SIVB, STI, STT, UNM, USB, WFC, ZION	31	94
						AIG, AMP, HIG, PFG, TMK	5	92,6						
BF-B, CHD, CPB, GIS, HRL, HSY, K, KO, MDLZ, MKC, MO, PEP, PG, PM	14	93,3	BF-B, CAG, CHD, CPB, GIS, HRL, HSY, K, KO, MDLZ, MKC, MO, PEP, PM, SJM	15	92,3	BF-B, HRL, HSY, KO, MKC, MO, PEP	7	94,3	CHD, KO, MDLZ, MKC, MO, PEP	6	91,4			
CAT, CMI, DE, EMR, ETN, FAST, GWW, PH, PNR, ROK, XYL	11	92,9				CMI, ETN, GWW, PH, PNR	5	92,2						
ADI, AMAT, AVGO, INTC, KLAC, LRCX, MCHP, MU, MXIM, NVDA, QRVO, SWKS, TXN, XLNX	14	92,9	ADI, AMAT, AVGO, INTC, KLAC, LRCX, MCHP, MU, MXIM, NVDA, QRVO, SWKS, TXN, XLNX	14	90,9				ADI, AMAT, AVGO, INTC, KLAC, LRCX, MCHP, MU, MXIM, NVDA, SWKS, TXN, XLNX	13	91,2			
						ARNC, DOV, FLR, FLS, JEC, PWR, URI	7	89,1						
ALXN, BIIB, CELG, INCY, REGN, VRTX	6	93,1	ALXN, AMGN, BIIB, CELG, GILD, ILMN, INCY, REGN, VRTX	9	94,1	ALXN, AMGN, BIIB, CELG, ILMN, INCY, REGN, VRTX	8	94,8	ALXN, AMGN, BIIB, CELG, ILMN, INCY, NKTR, REGN, VRTX	9	95,7	ALXN, AMGN, BIIB, INCY, REGN, VRTX	6	93,8
			ANTM, CI, CNC, HCA, HUM, LH, UHS, UNH, WCG	7	93	ANTM, CI, CNC, HCA, HUM, LH, UHS, UNH, WCG	9	92,4						

Figura 4.5. Clusters formati per k con valori $16 \leq k \leq 20$

4.5 Definizione parametro date

La definizione del parametro date è strettamente correlata a due domande fondamentali:

1. Qual è il periodo storico su cui vogliamo concentrare la nostra analisi?
2. Quanto deve essere lunga una serie temporale affinché il clustering dia dei buoni risultati?

La risposta a queste due domande ci dirà qual è il corretto range temporale da analizzare. Se però la risposta alla prima domanda è legata puramente ai requisiti della nostra applicazione, la seconda dipende strettamente da fattori tecnici. All'interno di questo lavoro esploreremo solo questa seconda domanda. La scelta della corretta lunghezza delle serie temporali finanziarie ai fini del clustering è stata approfondita all'interno del paper dal titolo "*Clustering Financial Time Series: How Long Is Enough?*"[33]. I ricercatori hanno utilizzato per il clustering di serie temporali lunghezze molto diverse tra loro: da 30 giorni a qualche anno[33]. L'importanza di definire correttamente la lunghezza delle serie temporali è cruciale nella buona riuscita dell'algoritmo di clustering. Se le serie risultassero troppo corte i cluster trovati sarebbero falsati, se troppo lunghe la loro dinamica potrebbe essere attenuata[33]. Gli esperimenti mostrano che la corretta dimensione delle serie temporali dipende strettamente dall'algoritmo di clustering utilizzato. Non è facile quindi definire, a priori, quale sia la corretta lunghezza da utilizzare. Per questo motivo in una prima fase proveremo ad applicare l'algoritmo con diverse lunghezze al fine di valutare i risultati ottenuti. Gli esperimenti sono stati eseguiti con i seguenti parametri "*date*":

- 1 mese: gennaio 2015;
- 3 mesi: da gennaio a marzo 2015;
- 6 mesi: da gennaio a giugno 2015;
- 1 anno: intero anno 2015;
- 2 anni: da gennaio 2015 a dicembre 2016.

Gli altri valori di input utilizzati sono descritti in Tabella 4.7

Parametri di input:	
n	100
k	15
date	a partire da gennaio 2015 range da 1 mese a 2 anni
somiglianza_threshold	85
percentuale	70
numero_minimo_titoli	5

Tabella 4.7. Parametri di input per definizione parametro *date*

In Figura 4.6 abbiamo i risultati dell'esperimento al variare di *date*. Come è possibile notare con l'aumentare del range temporale aumenta anche il numero di cluster formati. Questo comportamento è coerente infatti aumentando il tempo eventuali diversi comportamenti vengono mitigati.

1 mese			3 mesi			6 mesi			1 anno			2 anni		
Clusters	**N°**	**S.**	**Clusters**	**N°**	**S.**	**Clusters**	**N°**	**S.**	**Clusters**	**N°**	**S.**	**Clusters**	**N°**	**S.**
AEE, AEP, AWK, CMS, D, DTE, DUK, ED, EIX, ES, ETR, EXC, FE, LNT, NEE, NI, PCAR, PEG, PNW, PPL, SCG, SO, SRE, WEC, XEL	25	99,0	AEE, AEP, AES, AWK, CMS, CNP, D, DTE, DUK, ED, EIX, ES, ETR, EXC, FE, LNT, NEE, NI, PCG, PEG, PNW, PPL, SCG, SO, SRE, WEC, XEL	27	99, 8	AEE, AEP, AES, AWK, CMS, CNP, D, DTE, DUK, ED, EIX, ES, ETR, EXC, FE, LNT, NEE, NI, PCG, PEG, PNW, PPL, SCG, SO, SRE, WEC, XEL	27	99,56	AEE, AEP, AWK, CMS, CNP, D, DTE, DUK, ED, EIX, ES, ETR, EXC, FE, LNT, NEE, NI, PCG, PEG, PNW, PPL, SCG, SO, SRE, WEC, XEL	26	99,9	AEE, AEP, AWK, CMS, CNP, D, DTE, DUK, ED, EIX, ES, ETR, EXC, FE, LNT, NEE, NI, PCG, PEG, PNW, PPL, SCG, SO, SRE, WEC, XEL	26	99,7
AES, APA, APC, BHGE, COG, COP, CVX, CXO, DVN, EMN, EOG, FANG, FTI, HAL, HES, HP, KMI, LYB, MRO, NBL, NFX, NBL, NFX, NOV, OKE, OXY, PXD, SLB, TWTR, WMB, XEC, XOM	30	97, 1	APA, APC, BHGE, COG, COP, CVX, CXO, DVN, EOG, FANG, FTI, HAL, HES, HP, KMI, LYB, MRO, NBL, NFX, NOV, OKE, OXY, PXD, SLB, WMB, XEC, XOM	27	99, 8	APA, APC, BHGE, COG, COP, CVX, CXO, DVN, EOG, FANG, FTI, HAL, HES, HP, KMI, LYB, MRO, NBL, NFX, NOV, OKE, OXY, PXD, SLB, WMB, XEC, XOM	26	99,3	APA, APC, BHGE, COG, COP, CVX, CXO, DVN, EOG, FANG, FCX, FTI, HAL, HES, HP, KMI, LYB, MRO, NBL, NFX, NOV, NRG, OKE, OXY, PXD, SLB, WMB, XEC, XOM	29	96,6	APA, APC, BHGE, COG, COP, CVX, CXO, DVN, EOG, FANG, FTI, HAL, HES, HP, KMI, MRO, NBL, NFX, NOV, OKE, OXY, PXD, SLB, XEC, XOM	25	98,5
AIV, ARE, AVB, BXP, CBOE, DRE, EQR, ESS, EXR, FRT, HCP, KIM, MAA, MAC, MRK, PLD, REG, SLG, SPG, UDR, VNO, VTR, WELL	23	97, 9	AIV, ARE, AVB, BXP, DLR, DRE, EQR, ESS, EXR, FRT, HCP, KIM, MAA, O, PLD, PSA, REG, SLG, SPG, UDR, VNO, VTR, WELL	23	99, 5	AIV, AMT, ARE, AVB, BXP, DLR, DRE, EQR, ESS, EXR, FRT, HCP, HST, KIM, MAA, O, PLD, PSA, REG, SLG, SPG, UDR, VNO, VTR, WELL	25	98,7	AIV, ARE, AVB, BXP, DLR, DRE, EQR, ESS, EXR, FRT, HCP, IRM, KIM, MAA, O, PLD, PSA, REG, SLG, SPG, UDR, VNO, VTR, WELL	24	99,3	AIV, ARE, AVB, BXP, DLR, DRE, EQR, ESS, EXR, FRT, HCP, IRM, KIM, MAA, MAC, O, PLD, PSA, REG, SPG, UDR, VNO, VTR, WELL	24	98,2
BLK, C, CFG, CMA, HBAN, HIG, MET, MS, PNC, PRU, RJF, SCHW, STI, UNM, WFC, ZION	16	91, 2	AIG, BAC, BBT, BK, BLK, C, CFG, CMA, FITB, HBAN, HIG, JPM, KEY, LNC, MET, MS, MTB, NTRS, PBCT, PFG, PNC, RF, RJF, SCHW, SIVB, STI, STT, TROW, UNM, USB, WFC, ZION	32	94, 6	AIZ, BAC, BBT, BK, C, CFG, CMA, ETFC, FITB, HBAN, JPM, KEY, LNC, MET, MS, MTB, NTRS, PBCT, PNC, PRU, RF, RJF, SCHW, SIVB, STI, STT, UNM, USB, WFC, ZION	30	97	BAC, BBT, BK, C, CFG, CMA, COF, ETFC, FITB, GS, HBAN, JPM, KEY, LNC, MET, MS, MTB, NTRS, PBCT, PNC, PRU, RF, RJF, SCHW, SIVB, STI, STT, UNM, USB, WFC, ZION	31	98,1	AIG, AMP, AXP, BAC, BBT, BK, C, CFG, CMA, COF, DFS, ETFC, FITB, GS, HBAN, JPM, KEY, LNC, MET, MS, MTB, NTRS, PBCT, PFG, PNC, PRU, RF, RJF, SCHW, SIVB, STI, STT, SYF, TMK, UNM, USB, WFC, ZION	38	96,2
DHI, LEN, NKTR, PHM, WHR	5	89, 1												
FL, GPS, JWN, KSS, LB, M, WAT	7	93, 3				DG, DLTR, FL, GPS, JWN, M, ROST, TGT, TJX, ULTA	10	92,15	BBY, DG, DLTR, FL, GPS, JWN, KSS, LB, M, ROST, TGT, TJX, ULTA, VFC	14	96,2			
			AAL, ALK, DAL, LUV, UAL	5	95, 6	AAL, ALK, DAL, LUV, UAL	5	97,4	AAL, ALK, DAL, LUV, UAL	5	92,3	AAL, ALK, DAL, LUV, UAL	5	98
						CAT, EMR, ETN, JEC, PH, PNR, ROK, URI, XYL	9	91,2	CAT, CMI, DE, EMR, ETN, FAST, GWW, PH, PNR, ROK, XYL	11	94,5	CAT, CMI, DOV, EMR, ETN, FLS, JEC, PCAR	8	91,9
						CPB, GIS, KO, MO, PEP	5	90,5	BF-B, CAG, CHD, CL, CLX, CPB, GIS, HRL, HSY, K, KMB, KO, MDLZ, MKC, MO, PEP, PG, PM, SJM	19	95,5	BF-B, CAG, CHD, CL, CLX, CPB, GIS, HRL, HSY, K, KMB, KO, MKC, MO, PEP, PG, PM, SJM, SYY, TSN	20	96,5
						DOV, FCX, FLR, FLS, PWR	5	94,2						
						AJG, ALL, CB, CINF, MMC, PGR, RE, TRV	8	91,6	AJG, AON, BRK-B, CB, CINF, MMC, PGR, RE, TRV	9	90,9	ALL, CB, CINF, PGR, RE, TRV	6	96,8
												ABBV, AGN, ALXN, BIIB, BMY, CELG, GILD, INCY, REGN, VRTX	10	95
									ADI, AVGO, FFIV, INTC, KLAC, LRCX, MCHP, MU, MXIM, NVDA, QRVO, SWKS, TXN, XLNX	14	91,2	ADI, AMAT, AVGO, INTC, KLAC, LRCX, MCHP, MU, MXIM, NVDA, QRVO, SWKS, TXN, XLNX	14	94,7
												AMZN, FB, GOOG, GOOGL, NFLX	5	93,4
												BAX, BDX, BSX, COO, HOLX, ISRG, MDT, PKI, TMO, VAR, ZBH	11	91,8

Figura 4.6. Clusters formati per range temporali: 1 mese, 3 mesi, 6 mesi, 1 anno, 2 anni

4.6 Definizione parametri di similarità “somiglianza_threshold” e “percentuale”

Indaghiamo ora sui parametri “*somiglianza_threshold*” e “*percentuale*”. Come già anticipato questi parametri ci permettono di gestire il livello di somiglianza che i cluster devono avere. Sono parametri che entrano in gioco nella seconda fase della nostra pipeline. Infatti ci serviranno, dopo aver eseguito l’algoritmo di clustering n volte, per estrarre dalla Matrice di similarità i cluster ottenuti. Ci aspettiamo che configurando il parametro “*somiglianza_threshold*” con valori vicino ad n dovremmo generare cluster con similarità intra-cluster maggiore, viceversa con valori più vicino allo 0 dovremmo avere cluster con similarità intra-cluster minore. Il valore “*percentuale*” ci aspettiamo che abbia diretta influenza sul numero di oggetti all’interno del cluster. Valori vicini a 100 genereranno cluster più piccoli, al contrario valori vicini allo 0 genereranno cluster con più elementi. Anche il parametro “*somiglianza_threshold*” potrebbe però incidere sul numero di elementi all’interno del cluster. Infatti diminuendo il limite inferiore di somiglianza più oggetti faranno potenzialmente parte del cluster.

Per capire come questi due parametri influenzino la composizione dei cluster definiamo 3 livelli di similarità che corrispondono ai seguenti valori di “*somiglianza_threshold*” e “*percentuale*”:

- **Basso**: “*somiglianza_threshold*”=65 e “*percentuale*”=50;
- **Medio**: “*somiglianza_threshold*”=75 e “*percentuale*”=60;
- **Alto**: “*somiglianza_threshold*”=85 e “*percentuale*”=70;

L’algoritmo è stato quindi eseguito sugli stessi dati di input con le 3 configurazioni descritte e i parametri in Tabella 4.8

Parametri di input:	
n	100
k	15
date	2016
somiglianza_threshold	*basso*, *medio*, *alto*
percentuale	*basso*, *medio*, *alto*
numero_minimo_titoli	5

Tabella 4.8. Parametri di input per confronto livello di similarità “Alto”, “Medio” e “Basso”

In questo esperimento, a differenza dei precedenti, la definizione della matrice di similarità è avvenuta solo una volta. Il calcolo dei clusters è stato effettuato quindi con diversi parametri ma sulla stessa matrice. I risultati di questo esperimento sono mostrati in Figura 4.8. I dati sperimentali confermano le ipotesi che avevamo fatto. Con livello di similarità “Basso” il numero di elementi dei clusters è in media 20 con una similarità intra cluster di 84,4 mentre con livello di similarità “Alto” il numero di elementi dei clusters è in media 16 con una similarità intra cluster di 95,2. C’è da sottolineare come mediamente con i differenti parametri la composizione dei cluster cambia sempre ad eccezione di alcuni cluster che, a prescindere dai parametri, hanno componenti e livello di similarità uguale o molto simile. Interessante il caso del cluster formato da “AIV, ARE, AVB, CCI, DLR, DRE, EQR, ESS, EXR, FRT, HCP, IRM, KIM, MAA, MAC, O, PLD, PSA, REG, SPG, UDR, VTR, WELL” che nelle tre esecuzioni ha sempre gli stessi membri (e quindi lo stesso livello di similarità intra-cluster).

La sperimentazione ci mostra come sia determinante la definizione di questi due parametri. Anche sulla stessa matrice di similarità infatti non solo vengono estratti cluster con elementi differenti ma vengono “creati” o “eliminati” alcuni cluster. La scelta dei parametri “*somiglianza_threshold*” e “*percentuale*” da utilizzare è strettamente legata all’obiettivo dell’analisi.

Basso			Medio			Alto		
Clusters	**N°**	**S.**	**Clusters**	**N°**	**S.**	**Clusters**	**N°**	**S.**
A,ABMD,ABT,ALGN, BAX, BDX, BSX, CERN, CNC, COO, EW, HCA, HOLX, HSIC, IDXX, ILMN, IQV, ISRG, LH, MDT, MTD, PKI, RMD, SYK, TMO, UHS, VAR, WAT, WCG, XRAY, ZBH	31	84,5	A, ABMD, ALGN, BAX, BDX, BSX, CERN, CNC, COO, EW, HOLX, HSIC, IDXX, ILMN, IQV, ISRG, LH, MDT, MTD, PKI, RMD, SYK, TMO, UHS, VAR, WAT, XRAY, ZBH	28	86,8	ALGN, BAX, BDX, EW, HOLX, MDT, RMD, VAR, ZBH	9	95
AAL, ALK, DAL, HLT, LUV, MAR, NCLH, RCL, UAL	9	81,5	AAL, ALK, CCL, DAL, LUV, NCLH, RCL, UAL	8	87,6	AAL, ALK, DAL, LUV, UAL	5	98,8
ABBV, ABC, AGN, ALXN, AMGN, BIIB, BMY, CAH, CELG, CI, ESRX, GILD, HUM, INCY, LLY, MCK, MRK, MYL, NKTR, PFE, PRGO, REGN, VRTX	23	88,6	ABBV, ABC, AGN, ALXN, AMGN, BIIB, BMY, CAH, CELG, ESRX, GILD, INCY, LLY, MCK, MRK, MYL, NKTR, PFE, PRGO, REGN, VRTX	21	91,7	ABBV, AGN, ALXN, AMGN, BIIB, BMY, CELG, GILD, INCY, LLY, MCK, MRK, MYL, NKTR, PFE, PRGO, REGN, VRTX	18	92,9
ADI, AMAT, AMD, APH, AVGO, FTNT, INTC, IPGP, KEYS, KLAC, LRCX, MCHP, MU, MXIM, NVDA, QCOM, QRVO, SWKS, TXN, XLNX	20	80,6	ADI, AMAT, APH, AVGO, INTC, KLAC, LRCX, MCHP, MXIM, NVDA, QRVO, SWKS, TXN, XLNX	14	90,9	ADI, AMAT, AVGO, INTC, KLAC, LRCX, MCHP, MXIM, NVDA, QRVO, SWKS, TXN, XLNX	13	91,9
AEE, AEP, AES, AWK, CHD, CL, CLX, CMS, CNP, CPB, D, DTE, DUK, ED, EIX, ES, ETR, EXC, FE, GIS, HRL, HSY, K, KMB, KO, LNT, MKC, MO, NEE, NEM, NI, PCG, PEG, PEP, PG, PM, PNW, PPL, SCG, SJM, SO, SRE, SYY, T, TSN, VZ, WEC, XEL	48	84				AEE, AEP, AWK, CMS, CNP, D, DTE, DUK, ED, EIX, ES, ETR, EXC, FE, LNT, NEE, NI, PCG, PEG, PNW, PPL, SCG, SO, SRE, T, VZ, WEC, XEL	28	98,7
						CAG, CHD, CL, CLX, CPB, GIS, HRL, HSY, K, KMB, KO, MKC, MO, PEP, PG, PM, SJM, SYY, TSN	19	96
AIG, AMG, AMP, AXP, BAC, BBT, BEN, BK, BLK, C, CFG, CMA, COF, DFS, ETFC, FITB, GS, HBAN, IVZ, JPM, KEY, LNC, MET, MS, MTB, NTRS, PBCT, PFG, PNC, PRU, RF, RJF, SCHW, SIVB, STI, STT, SYF, TMK, TROW, UNM, USB, WFC, ZION	43	91,1	AIG, AMP, AXP, BAC, BBT, BK, C, CFG, CMA, COF, DFS, ETFC, FITB, GS, HBAN, IVZ, JPM, KEY, LNC, MET, MS, MTB, NTRS, PBCT, PFG, PNC, PRU, RF, RJF, SCHW, SIVB, STI, STT, SYF, TMK, TROW, UNM, USB, WFC, ZION	40	94,2	AIG, AMP, BAC, BBT, BK, C, CFG, CMA, COF, DFS, ETFC, FITB, GS, HBAN, IVZ, JPM, KEY, LNC, MET, MS, MTB, NTRS, PBCT, PFG, PNC, PRU, RF, RJF, SCHW, SIVB, STI, STT, SYF, TMK, UNM, USB, WFC, ZION	38	95,8
AIV, ARE, AVB, CCI, DLR, DRE, EQR, ESS, EXR, FRT, HCP, IRM, KIM, MAA, MAC, O, PLD, PSA, REG, SPG, UDR, VTR, WELL	23	97	AIV, ARE, AVB, CCI, DLR, DRE, EQR, ESS, EXR, FRT, HCP, IRM, KIM, MAA, MAC, O, PLD, PSA, REG, SPG, UDR, VTR, WELL	23	97	AIV, ARE, AVB, CCI, DLR, DRE, EQR, ESS, EXR, FRT, HCP, IRM, KIM, MAA, MAC, O, PLD, PSA, REG, SPG, UDR, VTR, WELL	23	97
ALLE, FBHS, MAS, MHK, SHW, SNA	6	76						
APA, APC, BHGE, COG, COP, CVX, CXO, DVN, EOG, FANG, FTI, HAL, HES, HP, KMI, MRO, NBL, NFX, NOV, NRG, OKE, OXY, PXD, SLB, XEC, XOM	26	95,9	APA, APC, BHGE, COG, COP, CVX, CXO, DVN, EOG, FANG, FTI, HAL, HES, HP, KMI, MRO, NBL, NFX, NOV, NRG, OKE, OXY, PXD, SLB, XEC, XOM	26	95,9	APA, APC, BHGE, COG, COP, CVX, CXO, DVN, EOG, FANG, FTI, HAL, HES, HP, KMI, MRO, NBL, NFX, NOV, OKE, OXY, PXD, SLB, XEC, XOM	25	97,35
CAT, CMI, DOV, EMR, ETN, FAST, FLS, GWW, JBHT, NUE, PCAR, PH, PNR, ROK	14	92,5	CAT, CMI, DOV, EMR, ETN, FAST, FLS, GWW, JBHT, NUE, PCAR, PH, PNR, ROK, TXT, XYL	16	90,6	CMI, DOV, EMR, ETN, NUE, PH, ROK	7	95,6
CBS, DIS, DISCA, DISCK, FOX, FOXA, VIAB	7	74,6						
CME, GD, HII, LLL, LMT, NOC, RTN	7	73,6						
DRI, FL, HBI, LB, ROST, TGT, TJX, TSCO	8	76,9						
			AFL, AIZ, AJG, ALL, AON, BRK-B, CB, CHRW, CINF, CMCSA, MMC, MMM, PGR, RE, TRV, UPS, WLTW	17	86,3	AJG, AON, BRK-B, CINF, MMC, RE	6	92,33
						AMZN, FB, GOOG, GOOGL, MSFT	5	91,9
						DRI, GPS, JWN, KSS, M, RL, TPR, VFC	8	95

Figura 4.7. Clusters formati per livelli di similarità “Basso”, “Medio”, “Alto”.

4.7 Clustering e similarità intra-cluster

In questa sezione esamineremo nel dettaglio una esecuzione dell'algoritmo. In particolare vedremo quanti oggetti vengono assegnati ai cluster rispetto al totale e rispetto ai GISC Sector, qual è la relazione tra i cluster creati e la tassonomia GISC, qual è la similarità intra-cluster.

In Tabella 4.9 vediamo i parametri con cui è stato eseguito l'algoritmo.

Parametri di input:	
n	100
k	15
date	2016
somiglianza_threshold	85
percentuale	70
numero_minimo_titoli	5

Tabella 4.9. Parametri di input per clustering e similarità intra-cluster

4.7.1 Numero di elementi clusterizzati

Fin dall'inizio di questo lavoro è stato sottolineato come la metodologia applicata renda possibile creare cluster con solo gli oggetti che realmente abbiano delle somiglianze tra loro. Al termine dell'algoritmo gli oggetti che faranno parte di un cluster saranno una parte rispetto al totale degli oggetti analizzati. L'obiettivo della metodologia è di clusterizzare solo gli oggetti che abbiano tra loro determinati livelli di similarità. La Figura 4.8 mostra come circa il 40% degli oggetti analizzati hanno fatto al termine parte di un cluster.

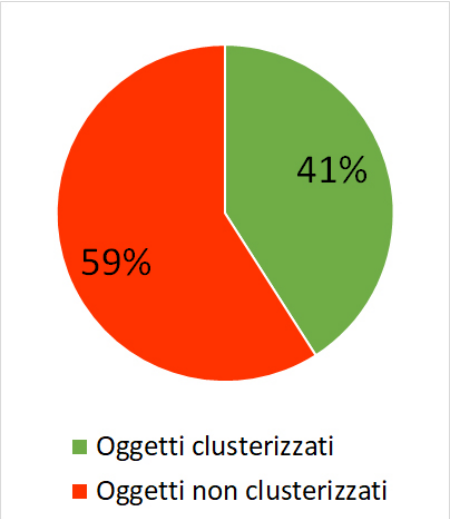

Figura 4.8. Numero di oggetti clusterizzati rispetto al totale

La Figura 4.9 invece ci mostra la percentuale di oggetti appartenenti ad un cluster (clusterizzati) e non appartenenti (non clusterizzati) suddivisi per GISC Sector. Questa visualizzazione ci consente di capire come i titoli clusterizzati appartengano maggiormente ad alcuni settori (Utilities, Energy, Real Estate, Financial) rispetto ad altri (Materials, Consumer Discretionary, Industrials, Communication Services). La motivazione di questa differenza tra titoli appartenenti a GISC Sector differenti potrebbe essere dovuta al fatto che in alcuni settori abbiamo degli andamenti maggiormente correlati tra loro rispetto al altri.

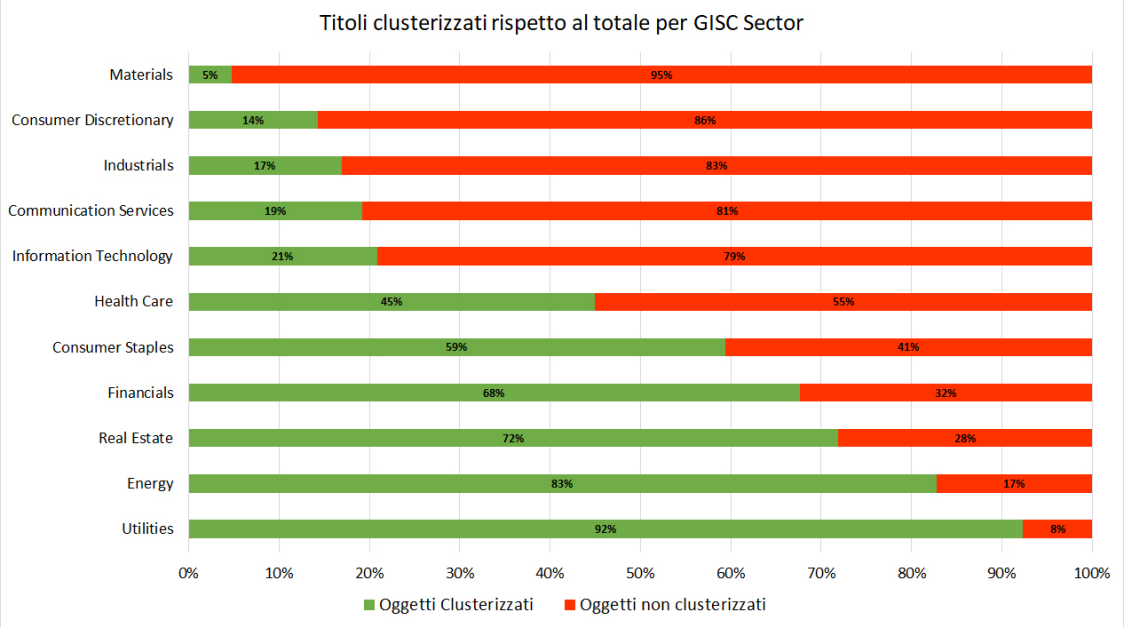

Figura 4.9. Numero di oggetti clusterizzati rispetto al totale per GISC Sector

4.7.2 Analisi cluster

I cluster formati dall'algoritmo sono 13 e sono visibili in Figura 4.10. Come possiamo vedere dai dati il numero di elementi di ogni cluster è molto variabile (da 5, il minimo consentito in base ai parametri di input, a 38). La similarità intra-cluster invece è per tutti i cluster molto alta e ha una media che supera i 95. Nella penultima e nell'ultima colonna è possibile vedere rispettivamente a quali GISC Industry (penultimo livello della tassonomia GISC) e GISC Sector (primo livello della tassonomia GISC) appartengono i titoli all'interno dei cluster. Interessante che si siano creati due cluster separati relativi ad uno stesso GISC Sector (sia per il settore Financials sia per il settore Health Care). Approfondiremo questa particolarità nella sezione 4.9. In generale tutti gli elementi dei cluster appartengono ad uno stesso GISC Sector ad eccezione dei cluster con ID 5, 10 e 12. I primi due hanno rispettivamente 2 e 1 elementi di un settore diverso rispetto a tutti gli altri (*T* e *VZ* per il cluster con ID 5 e *NUE* per il cluster con ID 10). All'interno del cluster con ID 12 invece, seppur i suoi elementi abbiano tassonomia GISC differente tra loro in quanto hanno core business differenti, troviamo tutte aziende che gravitano nel mondo dell'Information Technology e di Internet. Da segnalare inoltre che per alcuni cluster i livelli di similarità sono molto alti: ad esempio quasi 99 per il cluster con ID 2 che è composto da tutte le aziende con GISC Industry "Airlines" dello S&P500.

Nelle Figure 4.11 e 4.12 troviamo per ogni cluster la rappresentazione del suo prototype definito con il metodo Average (ogni punto è pari alla media dei rispettivi punti di tutti gli elementi del cluster). Nelle Figure 4.13 e 4.14 troviamo invece un dettaglio delle serie e del prototype di ogni cluster in un range temporale ristretto (gennaio-febbraio 2016).

Id	Clusters	Numero elementi	Similarità intra cluster	GISC Industry	GISC Sector
1	ALGN, BAX, BDX, EW, HOLX, MDT, RMD, VAR, ZBH	9	95,00	• Health Care Equipment & Supplies	• Health Care
2	AAL, ALK, DAL, LUV, UAL	5	98,80	• Airlines	• Industries
3	ABBV, AGN, ALXN, AMGN, BIIB, BMY, CELG, GILD, INCY, LLY, MCK, MRK, MYL, NKTR, PFE, PRGO, REGN, VRTX	18	92,90	• Pharmaceuticals • Biotechnology • Health Care Providers & Services	• Health Care
4	ADI, AMAT, AVGO, INTC, KLAC, LRCX, MCHP, MXIM, NVDA, QRVO, SWKS, TXN, XLNX	13	91,90	• Semiconductors & Semiconductor Equipment	• Information Technology
5	AEE, AEP, AWK, CMS, CNP, D, DTE, DUK, ED, EIX, ES, ETR, EXC, FE, LNT, NEE, NI, PCG, PEG, PNW, PPL, SCG, SO, SRE, T, VZ, WEC, XEL	28	98,70	• Multi-Utilities • Eletric Utilities • Water Utilities • Diversified Telecommunication Services	• Utilities • Communication Services
6	CAG, CHD, CL, CLX, CPB, GIS, HRL, HSY, K, KMB, KO, MKC, MO, PEP, PG, PM, SJM, SYY, TSN	19	96,00	• Food Product • Household Products • Beverages • Tobacco • Food & Staples Retailing • Personal Products	• Consumer Staples
7	AIG, AMP, BAC, BBT, BK, C, CFG, CMA, COF, DFS, ETFC, FITB, GS, HBAN, IVZ, JPM, KEY, LNC, MET, MS, MTB, NTRS, PBCT, PFG, PNC, PRU, RF, RJF, SCHW, SIVB, STI, STT, SYF, TMK, UNM, USB, WFC, ZION	38	95,80	• Insurance • Capital Markets • Banks • Consumer Finance • Thrifts & Mortgage Finance	• Financials
8	AIV, ARE, AVB, CCI, DLR, DRE, EQR, ESS, EXR, FRT, HCP, IRM, KIM, MAA, MAC, O, PLD, PSA, REG, SPG, UDR, VTR, WELL	23	97,00	• Equity Real Estate Investment Trusts (REITs)	• Real Estate
9	APA, APC, BHGE, COG, COP, CVX, CXO, DVN, EOG, FANG, FTI, HAL, HES, HP, KMI, MRO, NBL, NFX, NOV, OKE, OXY, PXD, SLB, XEC, XOM	25	97,35	• Energy Equipment & Services • Oil, Gas & Consumable Fuels	• Energy
10	CMI, DOV, EMR, ETN, NUE, PH, ROK	7	95,60	• Machinery • Electrical Equipment • Metals & Mining	• Industrials • Materials
11	AJG, AON, BRK-B, CINF, MMC, RE	6	92,33	• Insurance • Diversified Financial Services	• Financials
12	AMZN, FB, GOOG, GOOGL, MSFT	5	91,90	• Internet & Direct Marketing Retail • Interactive Media & Services • Software	• Consumer Discretionary • Communication Services • Information Technology
13	DRI, GPS, JWN, KSS, M, RL, TPR, VFC	8	95,00	• Hotels, Restaurants & Leisure • Specialty Retail • Multiline Retail • Textiles, Apparel & luxury goods	• Consumer Discretionary

Figura 4.10. Clusters formati 2016

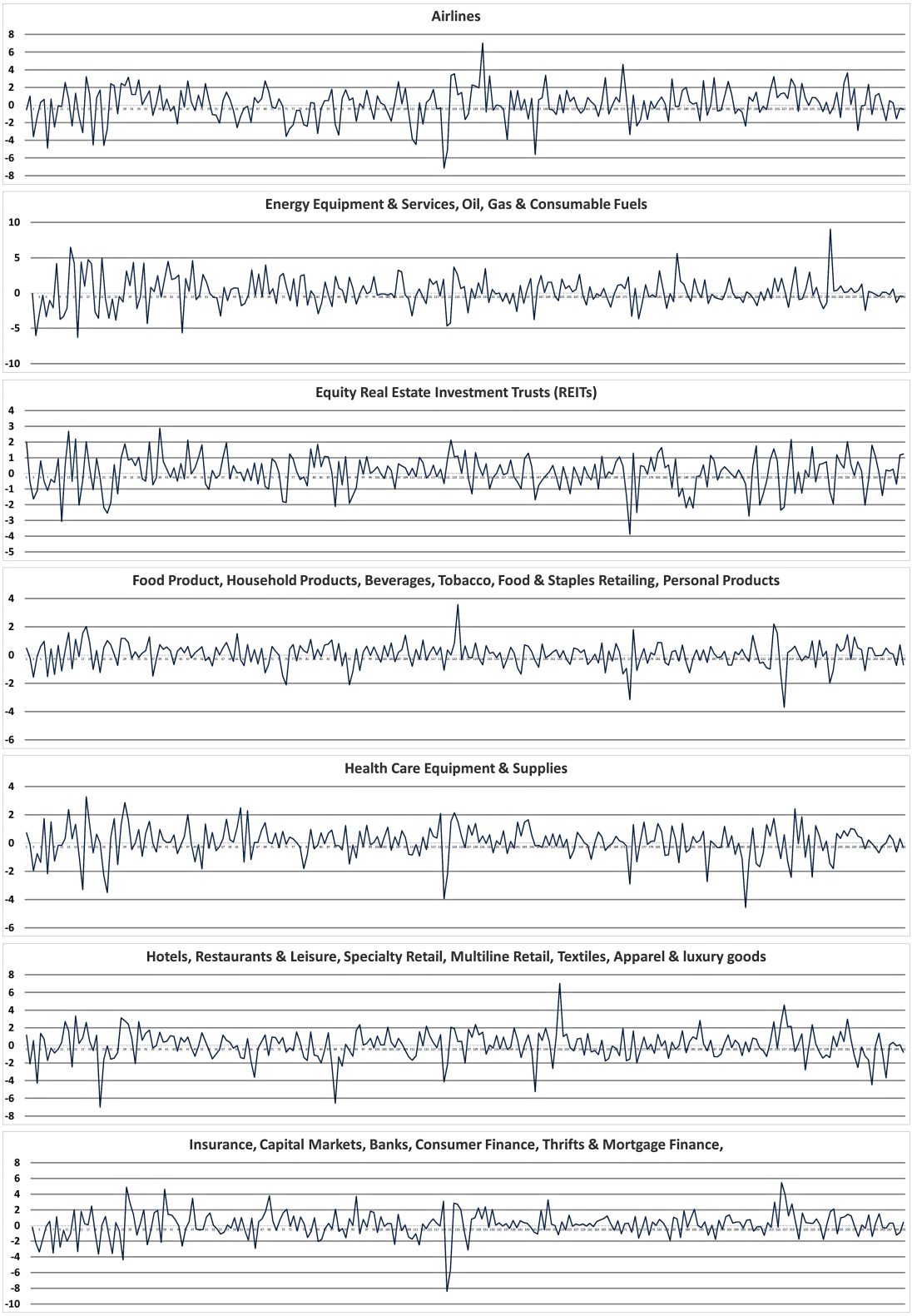

Figura 4.11. Media serie temporali clusters 2016 (1)

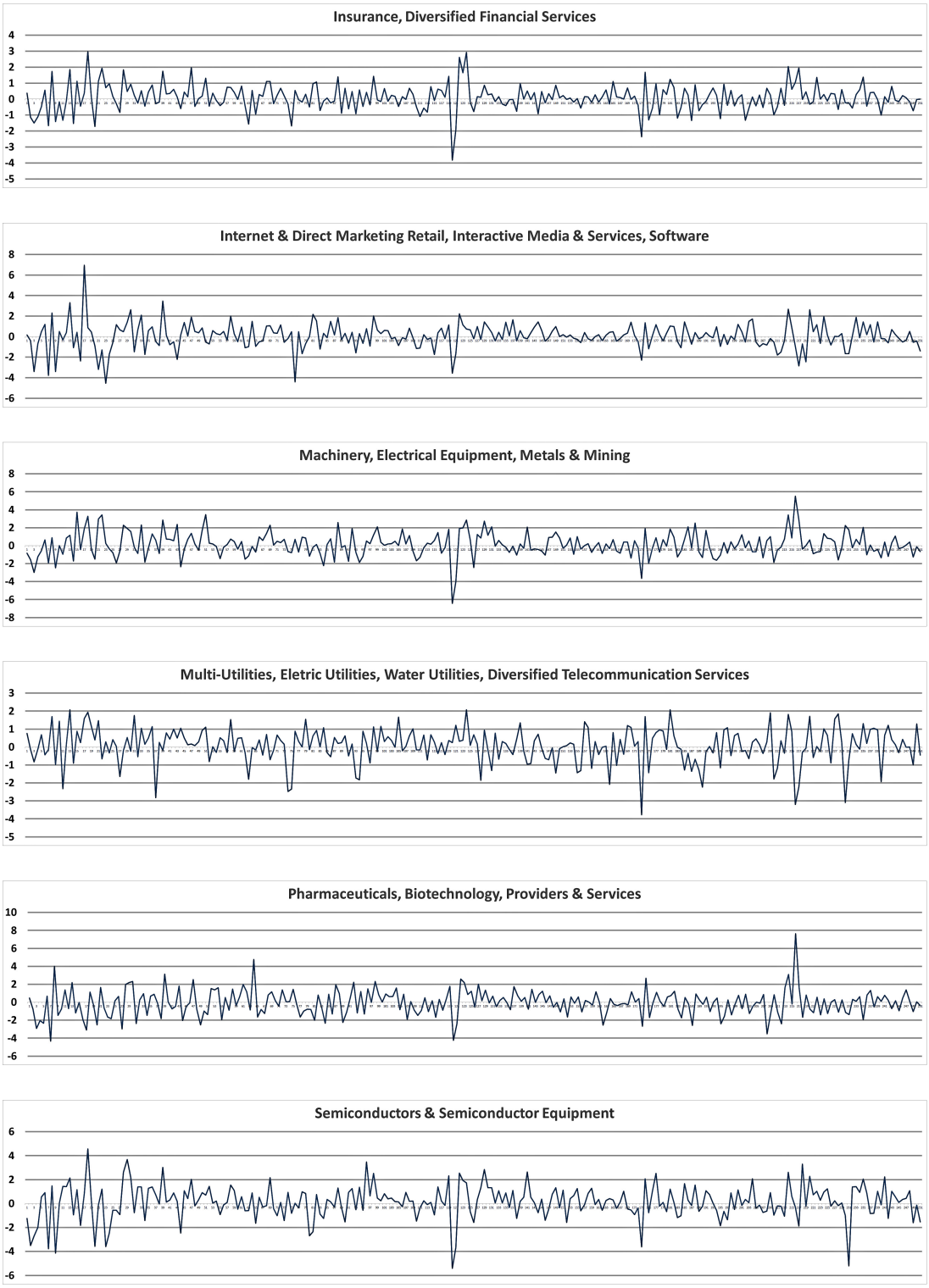

Figura 4.12. Media serie temporali clusters 2016 (2)

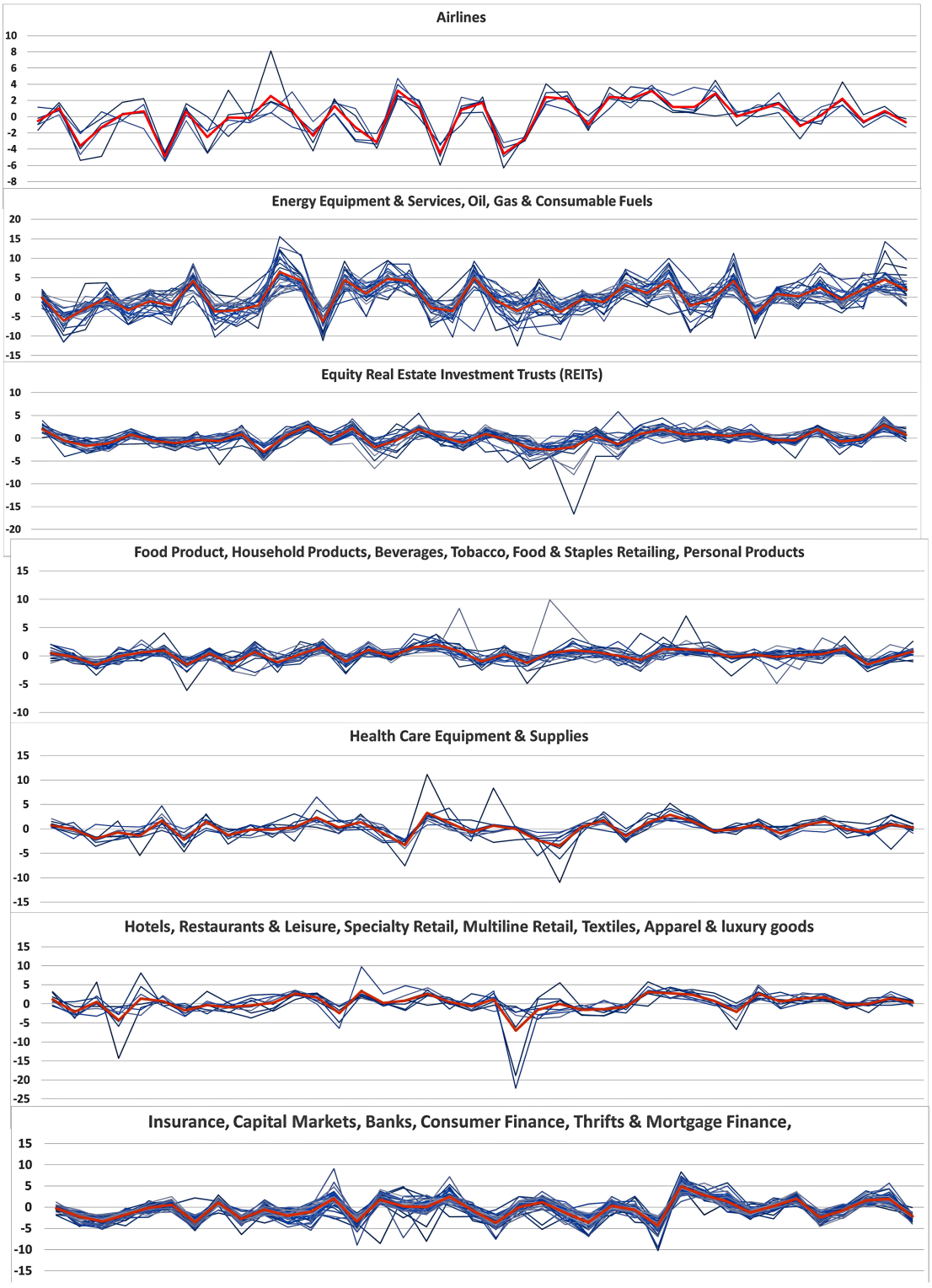

Figura 4.13. Dettaglio serie temporali clusters Gennaio-Febbraio 2016 (1)

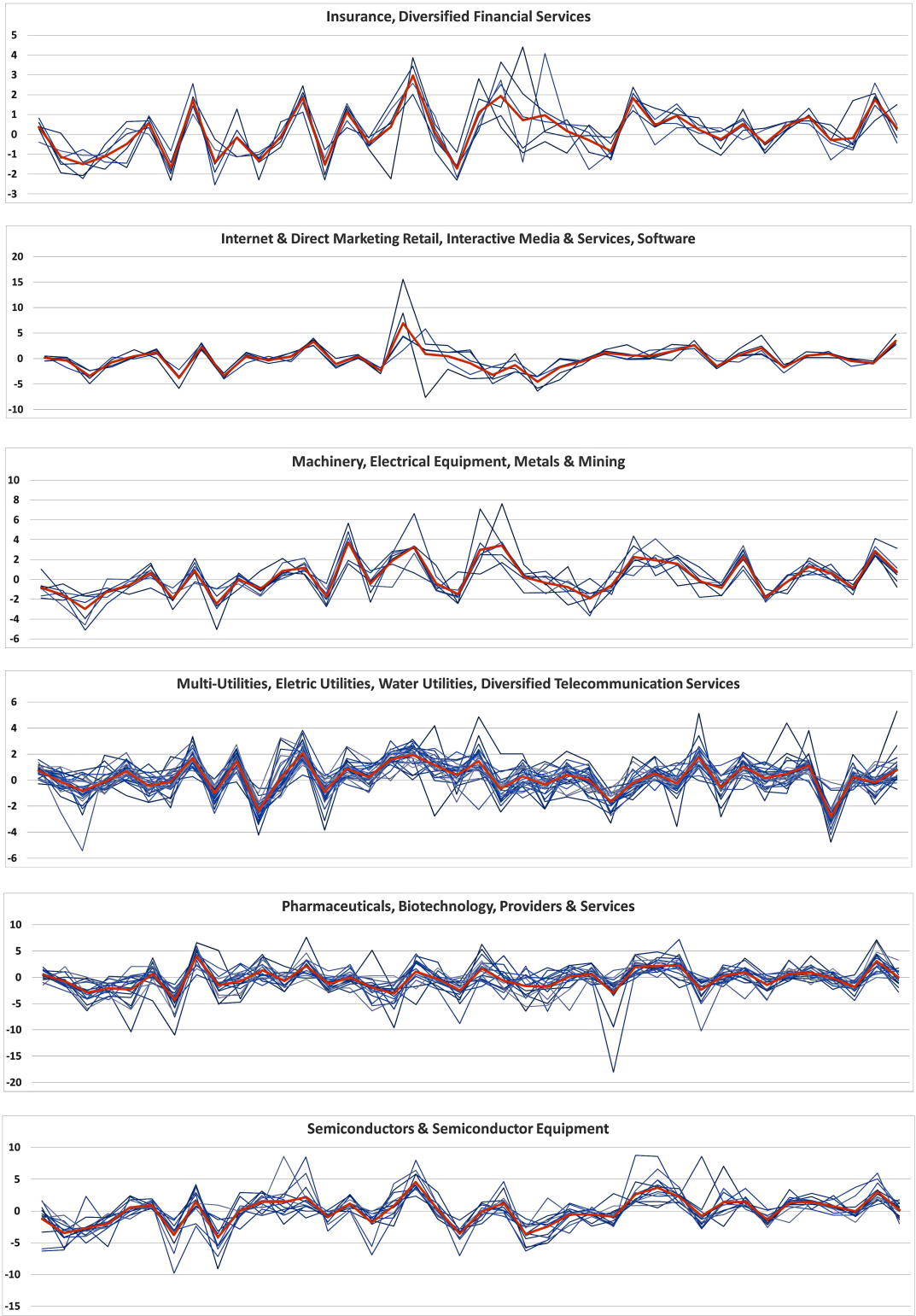

Figura 4.14. Dettaglio serie temporali clusters Gennaio-Febbraio 2016 (2)

4.8 Similarità tra singoli oggetti

Un altro importante risultato di questo lavoro è poter analizzare la distanza tra singoli titoli. Questo risultato si ottiene andando a vedere all'interno della matrice di similarità il numero di volte che un titolo è stato assegnato allo stesso cluster di un altro. Vedremo adesso alcuni esempi. Ognuno avrà una dicitura contenente il titolo che si sta osservando e le seguenti informazioni:

- **SIMBOLO** (Nome azienda - GISC Sector - GISC Sub Industry)

- **AMZN** (Amazon - Consumer Discretionary - Internet & Direct Marketing Retail)

  Ordinando in modo decrescente la colonna della matrice di similarità relativa al titolo *AMZN* ed estraendo le prime 10 righe abbiamo il risultato in Tabella 4.10. Come possiamo vedere i titoli più vicini sono tutti di aziende in ambito Information Technology.

Simbolo	Nome	GISC Sector	GICS Sub Industry	Somiglianza
GOOG	Alphabet Inc Class C	Communication Services	Interactive Media & Services	95
FB	Facebook, Inc.	Communication Services	Interactive Media & Services	92
MSFT	Microsoft Corp.	Information Technology	Systems Software	89
ATVI	Activision Blizzard	Communication Services	Interactive Home Entertainment	81
NFLX	Netflix Inc.	Communication Services	Movies & Entertainment	80
EA	Electronic Arts	Communication Services	Interactive Home Entertainment	78
SBUX	Starbucks Corp.	Consumer Discretionary	Restaurants	76
PYPL	PayPal	Information Technology	Data Processing & Outsourced Services	76
EXPE	Expedia Group	Consumer Discretionary	Internet & Direct Marketing Retail	70
ADBE	Adobe Systems Inc	Information Technology	Application Software	70

Tabella 4.10. Top 10 titoli più vicini ad AMZN (Amazon)

Una possibile visualizzazione grafica dei risultati potrebbe essere quella in Figura 4.15.

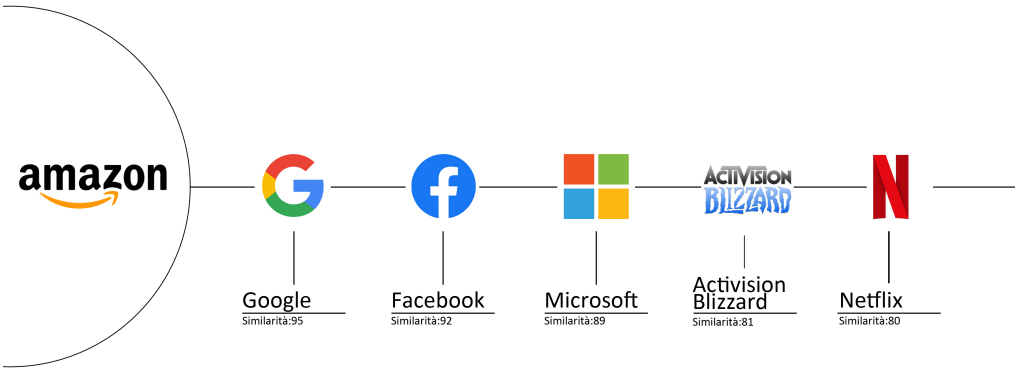

Figura 4.15. Top 5 titoli più vicini al titolo AMZN (Amazon)

- **AAL** (American Airlines Group - Industrials - Airlines) Interessante il caso di AAL (American Airlines) in cui le aziende più vicine e con somiglianza molto alta sono quelle della sua stessa GISC Sub Industry e cioè "Airlines" con cui, tra l'altro, forma anche il cluster visto precedentemente. Quelle subito dopo sono tutte aziende vicine al mondo dei viaggi per ovvi motivi strettamente connesse alle prime. In Tabella 4.11 il dettaglio.

Simbolo	Nome	GISC Sector	GICS Sub Industry	Somiglianza
DAL	Delta Air Lines Inc.	Industrials	Airlines	100
ALK	Alaska Air Group Inc	Industrials	Airlines	99
UAL	United Airlines Holdings	Industrials	Airlines	99
LUV	Southwest Airlines	Industrials	Airlines	98
RCL	Royal Caribbean Cruises Ltd	Consumer Discretionary	Hotels, Resorts & Cruise Lines	80
NCLH	Norwegian Cruise Line Holdings	Consumer Discretionary	Hotels, Resorts & Cruise Lines	79
CCL	Carnival Corp.	Consumer Discretionary	Hotels, Resorts & Cruise Lines	78
MAR	Marriott Int'l.	Consumer Discretionary	Hotels, Resorts & Cruise Lines	71
HLT	Hilton Worldwide Holdings Inc	Consumer Discretionary	Hotels, Resorts & Cruise Lines	68
APTV	Aptiv Plc	Consumer Discretionary	Auto Parts & Equipment	62

Tabella 4.11. Top 10 titoli più vicini ad AAL (American Airlines)

- **XEL** (Xcel Energy Inc - Utilities - Multi-Utilities)

  Per analizzare il caso di *XEL*, e di tutti i titoli appartenenti al cluster con ID 5 ("Utilities") l'analisi si amplia ai 30 titoli più vicini. Questo perché i titoli del settore Utilities hanno tra loro una elevatissima similarità: i 25 titoli più vicini a *XEL* hanno con il titolo un livello si similarità di 100 su 100. In Tabella 4.12 il dettaglio.

Simbolo	Nome	GISC Sector	GICS Sub Industry	Somiglianza
AEP	American Electric Power	Utilities	Electric Utilities	100
AWK	American Water Works Company Inc	Utilities	Water Utilities	100
CMS	CMS Energy	Utilities	Multi-Utilities	100
CNP	CenterPoint Energy	Utilities	Multi-Utilities	100
DTE	DTE Energy Co.	Utilities	Multi-Utilities	100
DUK	Duke Energy	Utilities	Electric Utilities	100
D	Dominion Energy	Utilities	Electric Utilities	100
ED	Consolidated Edison	Utilities	Electric Utilities	100
EIX	Edison Int'l	Utilities	Electric Utilities	100
ES	Eversource Energy	Utilities	Multi-Utilities	100
ETR	Entergy Corp.	Utilities	Electric Utilities	100
EXC	Exelon Corp.	Utilities	Multi-Utilities	100
FE	FirstEnergy Corp	Utilities	Electric Utilities	100
LNT	Alliant Energy Corp	Utilities	Electric Utilities	100
NEE	NextEra Energy	Utilities	Multi-Utilities	100
NI	NiSource Inc.	Utilities	Multi-Utilities	100
PCG	Rimossa	Rimossa	Rimossa	100
PEG	Public Serv. Enterprise Inc.	Utilities	Electric Utilities	100
PNW	Pinnacle West Capital	Utilities	Multi-Utilities	100
PPL	PPL Corp.	Utilities	Electric Utilities	100
SCG	Rimossa	Rimossa	Rimossa	100
SO	Southern Co.	Utilities	Electric Utilities	100
SRE	Sempra Energy	Utilities	Multi-Utilities	100
WEC	Wec Energy Group Inc	Utilities	Electric Utilities	100
XEL	Xcel Energy Inc	Utilities	Multi-Utilities	100
T	AT&T Inc.	Communication Services	Integrated Telecommunication Services	92
VZ	Verizon Communications	Communication Services	Integrated Telecommunication Services	89
AES	AES Corp	Utilities	Independent Power Producers & Energy Traders	84
NEM	Newmont Goldcorp	Materials	Gold	75
CLX	The Clorox Company	Consumer Staples	Household Products	74

Tabella 4.12. Top 30 titoli più vicini a XEL (Xcel Energy)

- **APA** (Apache Corporation - Energy - Oil & Gas Exploration & Production)

  Analizziamo ora i titoli più vicini ad *APA*, membro del cluster con ID 9 ("Energy"). Anche gli elementi più vicini a questo titolo hanno una similarità molto elevata. In particolare tutti quelli con similarità uguale a 100 hanno la stessa GISC Sub Industry di *APA* cioè "Oil & Gas Exploration & Production". In Tabella 4.13 il dettaglio.

Simbolo	Nome	GISC Sector	GICS Sub Industry	Somiglianza
APC	Anadarko Petroleum Corp	Energy	Oil & Gas Exploration & Production	100
COP	ConocoPhillips	Energy	Oil & Gas Exploration & Production	100
CXO	Concho Resources	Energy	Oil & Gas Exploration & Production	100
DVN	Devon Energy	Energy	Oil & Gas Exploration & Production	100
EOG	EOG Resources	Energy	Oil & Gas Exploration & Production	100
FANG	Diamondback Energy	Energy	Oil & Gas Exploration & Production	100
MRO	Marathon Oil Corp.	Energy	Oil & Gas Exploration & Production	100
NBL	Noble Energy Inc	Energy	Oil & Gas Exploration & Production	100
PXD	Pioneer Natural Resources	Energy	Oil & Gas Exploration & Production	100
XEC	Cimarex Energy	Energy	Oil & Gas Exploration & Production	100
OXY	Occidental Petroleum	Energy	Oil & Gas Exploration & Production	99
HP	Helmerich & Payne	Energy	Oil & Gas Drilling	98
HAL	Halliburton Co.	Energy	Oil & Gas Equipment & Services	98
SLB	Schlumberger Ltd.	Energy	Oil & Gas Equipment & Services	98
CVX	Chevron Corp.	Energy	Integrated Oil & Gas	97
BHGE	Baker Hughes, a GE Company	Energy	Oil & Gas Equipment & Services	97
COG	Cabot Oil & Gas	Energy	Oil & Gas Exploration & Production	97
XOM	Exxon Mobil Corp.	Energy	Integrated Oil & Gas	96
FTI	TechnipFMC	Energy	Oil & Gas Equipment & Services	96
NOV	National Oilwell Varco Inc.	Energy	Oil & Gas Equipment & Services	94
KMI	Kinder Morgan	Energy	Oil & Gas Storage & Transportation	94
OKE	ONEOK	Energy	Oil & Gas Storage & Transportation	94
NRG	NRG Energy	Utilities	Independent Power Producers & Energy Traders	79
WMB	Williams Cos.	Energy	Oil & Gas Storage & Transportation	35
FCX	Freeport-McMoRan Inc.	Materials	Copper	26

Tabella 4.13. Top 25 titoli più vicini a APA (Apache Corporation)

- **CMCSA** (Comcast Corp. - Communication Services - Cable & Satellite)

  Diverso il caso del titolo *CMCSA* che, seppur con somiglianze non altissime, ha vicino titoli appartenenti ai GISC Sector "Financials" e "Industrials". In Tabella 4.14 il dettaglio.

Simbolo	Nome	GISC Sector	GICS Sub Industry	Somiglianza
MMC	Marsh & McLennan	Financials	Insurance Brokers	90
AON	Aon plc	Financials	Insurance Brokers	89
CINF	Cincinnati Financial	Financials	Property & Casualty Insurance	89
UPS	United Parcel Service	Industrials	Air Freight & Logistics	88
BRK-B	Berkshire Hathaway	Financials	Multi-Sector Holdings	88
AJG	Arthur J. Gallagher & Co.	Financials	Insurance Brokers	87
RE	Everest Re Group Ltd.	Financials	Reinsurance	87
CHRW	C. H. Robinson Worldwide	Industrials	Air Freight & Logistics	86
WLTW	Willis Towers Watson	Financials	Insurance Brokers	84
ALL	Allstate Corp	Financials	Property & Casualty Insurance	84

Tabella 4.14. Top 10 titoli più vicini a CMCSA (Comcast)

4.9 Drill down su sector o cluster

Per l'analisi di tipo "Drill down", già descritta nella sezione 3.2.8, è stato deciso di analizzare gli oggetti partendo dalla matrice di similarità creata con l'analisi della sezione 4.9 ed eseguendo 3 volte l'estrazione dei cluster, come già descritto nella sezione 3.2.5, con 3 diversi parametri di similarità:

- **1° step**: "*somiglianza_threshold*"=65 e "*percentuale*"=50;
- **2° step**: "*somiglianza_threshold*"=85 e "*percentuale*"=50;
- **3° step**: "*somiglianza_threshold*"=95 e "*percentuale*"=50;

La scelta di questi parametri ha l'obiettivo di creare per ogni step successivo dei cluster a più alta similarità interna (infatti il valore di "*somiglianza_threshold*" è crescente da 65 a 95) ma creando cluster numerosi (infatti il valore "*percentuale*" viene mantenuto costante a 50).

In Figura 4.16 vediamo i risultati dell'applicazione di questa metodologia a tutti i titoli assegnati dalla tassonomia GISC al sector "Health Care". Gli oggetti analizzati sono 60. Elenchiamo i simboli per completezza: A, ABBV, ABC, ABMD, ABT, AGN, ALGN, ALXN, AMGN, ANTM, BAX, BDX, BIIB, BMY, BSX, CAH, CELG, CERN, CI, CNC, COO, CVS, DHR, DVA, EW, GILD, HCA, HOLX, HSIC, HUM, IDXX, ILMN, INCY, IQV, ISRG, JNJ, LH, LLY, MCK, MDT, MRK, MTD, MYL, NKTR, PFE, PKI, PRGO, REGN, RMD, SYK, TMO, UHS, UNH, VAR, VRTX, WAT, WCG, XRAY, ZBH, ZTS.

1° step			2° step			3° step		
Clusters	**N°**	**S.**	**Clusters**	**N°**	**S.**	**Clusters**	**N°**	**S.**
A, ABMD, ABT, ALGN, BAX, BDX, BSX, CERN, CNC, COO, EW, HCA, HOLX, HSIC, IDXX, ILMN, IQV, ISRG, LH, MDT, MTD, PKI, RMD, SYK, TMO, UHS, VAR, WAT, WCG, XRAY, ZBH	31	80,56	A, MTD, PKI, TMO, WAT	5	91,80			
			ABMD, ALGN, BAX, BDX, BSX, COO, EW, HOLX, IDXX, ISRG, MDT, RMD, SYK, VAR, XRAY, ZBH	16	94,86	BAX, BDX, BSX, COO, EW, HOLX, MDT, SYK, XRAY, ZBH	10	97,26
			ABT, ILMN, IQV	3	83,00			
ABBV, ABC, AGN, ALXN, AMGN, ANTM, BIIB, BMY, CAH, CELG, CI, GILD, HUM, INCY, LLY, MCK, MRK, MYL, NKTR, PFE, PRGO, REGN, VRTX	23	88,92	ABBV, ABC, AGN, ALXN, AMGN, BIIB, BMY, CAH, CELG, GILD, INCY, LLY, MCK, MRK, MYL, NKTR, PFE, PRGO, REGN, VRTX	20	93,66	ABBV, ALXN, AMGN, BIIB, BMY, CELG, GILD, INCY, MYL, NKTR, REGN, VRTX	12	97,63
						ABC, CAH, MCK	3	98,00
			ANTM, CI, HUM	3	92,00			

Figura 4.16. Clusters per drill down su GISC Sector "Health Care"

Per questo genere di analisi è stato ideato un nuovo metodo di visualization, che vediamo in Figura 4.17, con l'obiettivo di rendere questi dati facilmente fruibili da qualsiasi osservatore. Questo metodo descrive i cluster con circonferenze di grandezza ed intensità di colore variabile. La grandezza delle circonferenze è variabile a seconda della somma del market cap degli elementi del cluster. Il colore descrive invece a quale step dell'analisi ci stiamo riferendo e ha una intensità diversa a seconda della similarità interna del cluster (intensità crescente indica similarità interna maggiore). I diversi colori, come detto, indicano i differenti step di analisi, in particolare: (a) in azzurro racchiudiamo tutti i simboli oggetto di analisi (in questo caso tutti i simboli del GISC Sector "Health Care"); (b) in verde abbiamo i clusters formati dal 1° step di analisi e i singoli titoli che non sono stati assegnati a nessun cluster; (c) in rosso abbiamo i cluster del 2° step; (d) infine in blu troviamo i cluster del 3° step. Per ogni cluster troviamo in grassetto un elenco delle "GISC Industry" dei titoli che formano il cluster, la similarità interna del cluster, la somma del market cap dei titoli che ne fanno parte e il numero di elementi.

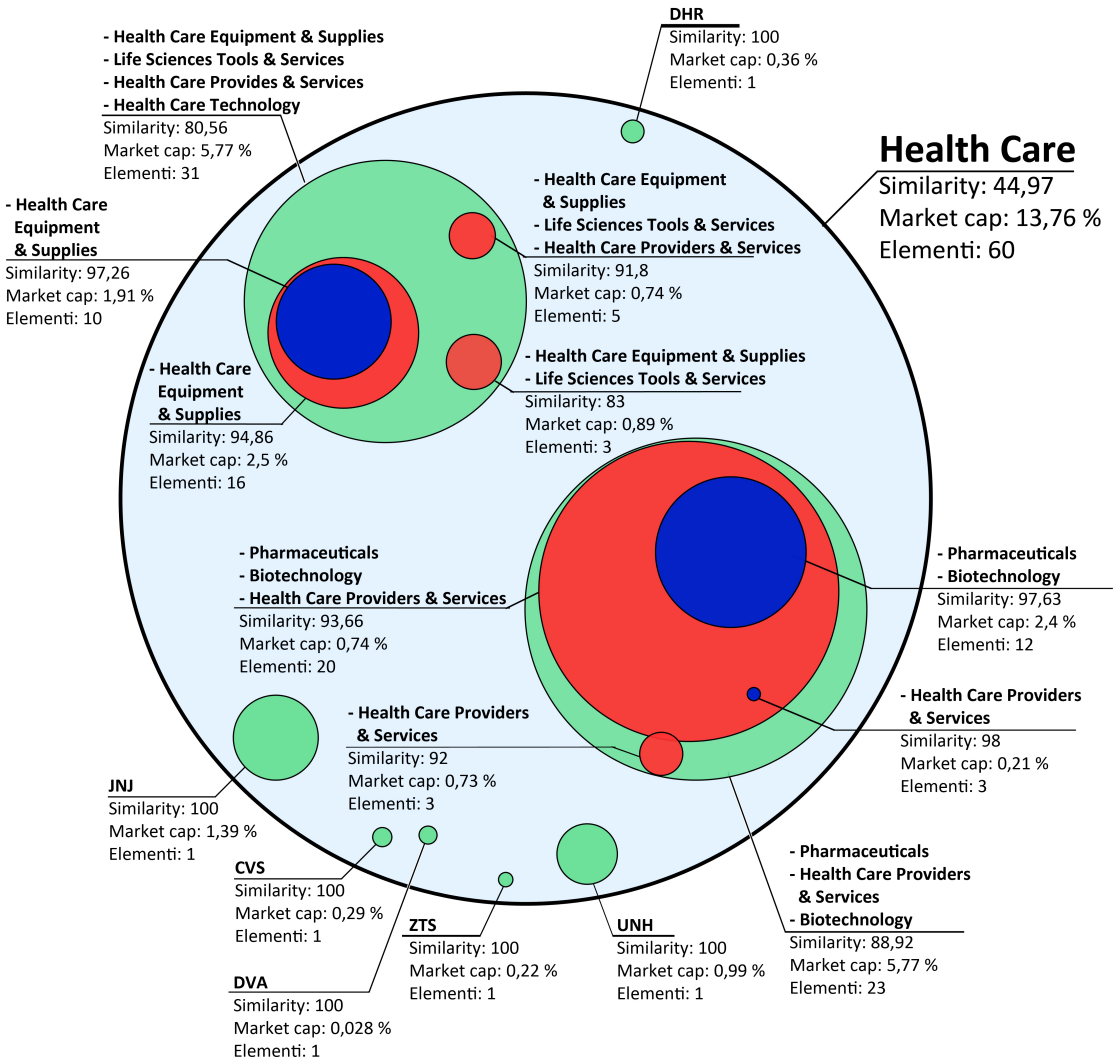

Figura 4.17. Visualization drill down su GISC Sector "Health Care"

Vene riportata per completezza in Figura 4.18 la porzione di matrice di somiglianza analizzata con evidenziate le similarità più alte.

Figura 4.18. Matrice di somiglianza drill down su GISC Sector “Health Care”

Al fine di validare questo metodo lo stesso genere di analisi è stata effettuata anche su altri titoli. Vediamo adesso i risultati ottenuti sui titoli appartenenti al GISC Sector "Energy". In Figura 4.19 abbiamo i risultati dei 3 step, mentre in Figura 4.20 abbiamo applicato a questi dati lo stesso metodo di visualization descritto in precedenza.

Gli oggetti analizzati sono 65. Elenchiamo i simboli per completezza: AFL, AIG, AIZ, AJG, ALL, AMG, AMP, AON, AXP, BAC, BBT, BEN, BK, BLK, BRK-B, C, CB, CBOE, CFG, CINF, CMA, CME, COF, DFS, ETFC, FITB, GS, HBAN, HIG, ICE, IVZ, JPM, KEY, L, LNC, MCO, MET, MMC, MS, MSCI, MTB, NDAQ, NTRS, PBCT, PFG, PGR, PNC, PRU, RE, RF, RJF, SCHW, SIVB, SPGI, STI, STT, SYF, TMK, TROW, TRV, UNM, USB, WFC, WLTW, ZION.

1° step			2° step			3° step		
Clusters	**N°**	**S.**	**Clusters**	**N°**	**S.**	**Clusters**	**N°**	**S.**
AFL, AIZ, AJG, ALL, AON, BRK-B, CB, CINF, HIG, MMC, PGR, RE, TRV, WLTW	14	84,59	AIZ, AJG, ALL, BRK-B, CB, CINF, MMC, PGR, RE, TRV	10	91,60	ALL, CB, CINF, PGR, RE, TRV,	7	97,06
AIG, AMG, AMP, AXP, BAC, BBT, BEN, BK, BLK, C, CFG, CMA, COF, DFS, ETFC, FITB, GS, HBAN, IVZ, JPM, KEY, LNC, MET, MS, MTB, NTRS, PBCT, PFG, PNC, PRU, RF, RJF, SCHW, SIVB, STI, STT, SYF, TMK, TROW, UNM, USB, WFC, ZION	43	91,44	AIG, AMP, AXP, BAC, BBT, BK, C, CFG, CMA, COF, DFS, ETFC, FITB, GS, HBAN, IVZ, JPM, KEY, LNC, MET, MS, MTB, NTRS, PBCT, PFG, PNC, PRU, RF, RJF, SCHW, SIVB, STI, STT, SYF, TMK, UNM, USB, WFC, ZION	39	95,84	BAC, BBT, BK, C, CFG, CMA, COF, DFS, ETFC, FITB, GS, HBAN, JPM, KEY, MS, MTB, NTRS, PBCT, PNC, RF, RJF, SCHW, SIVB, STI, STT, SYF, USB, WFC, ZION	29	97,96
CBOE, ICE, NDAQ	3	75,3						

Figura 4.19. Clusters per drill down su GISC Sector "Financial"

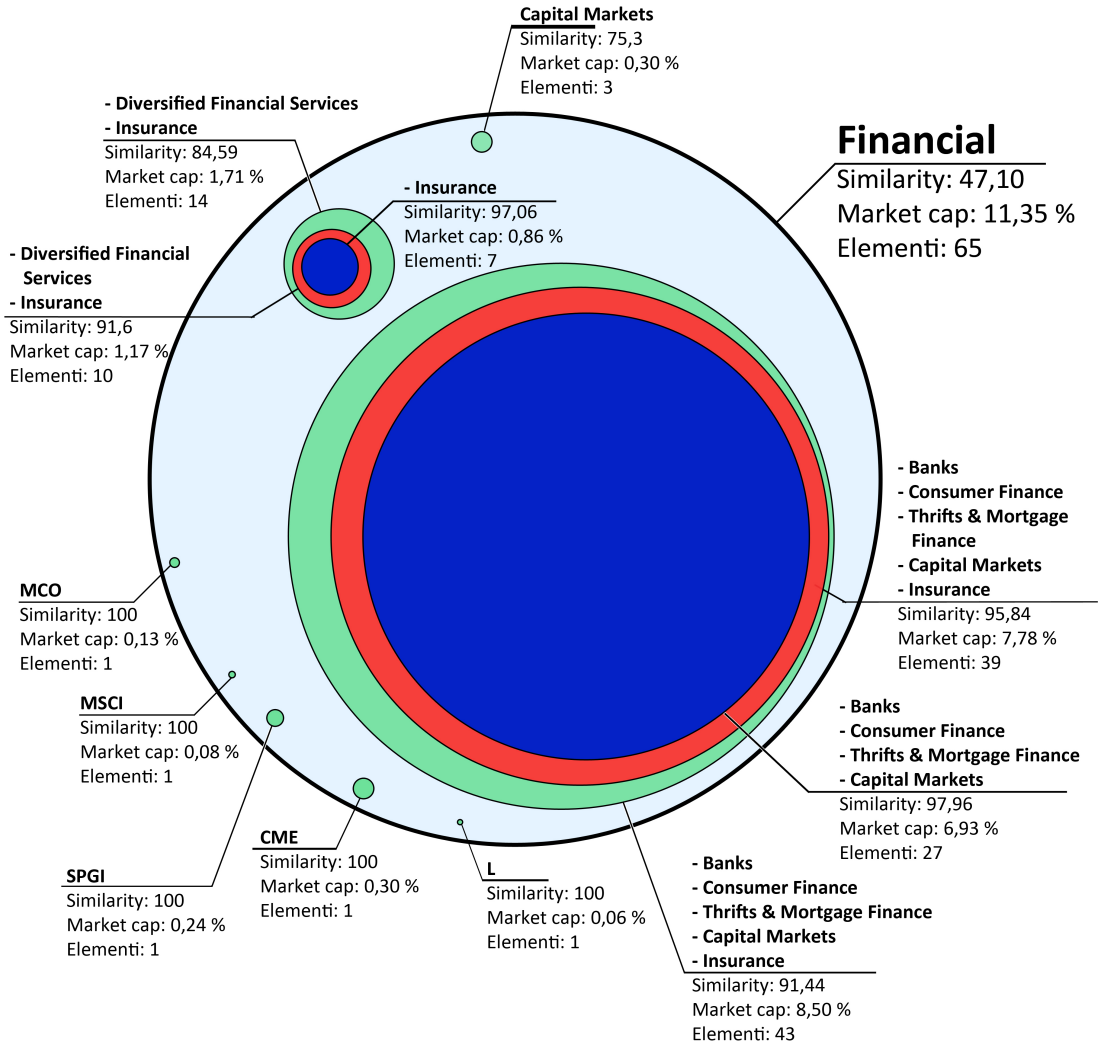

Figura 4.20. Visualization drill down su GISC Sector “Financial”

Vene riportata per completezza in Figura 4.21 la porzione di matrice di somiglianza analizzata con evidenziate le similarità più alte.

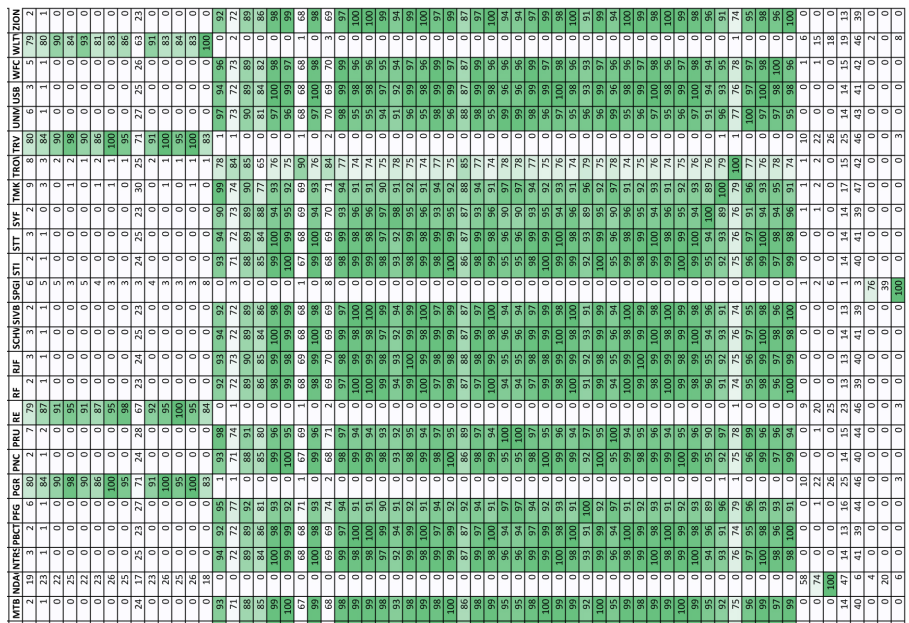

Figura 4.21. Matrice di somiglianza drill down su GISC Sector “Financial”

Capitolo 5

Conclusioni e sviluppi futuri

5.1 Conclusioni

La metodologia applicata in questo lavoro ha dimostrato di poter, in modo efficiente, estrarre da un dataset di serie temporali finanziarie le relazioni presenti tra i singoli titoli o tra gruppi di titoli (cluster). Le analisi sono state effettuate con differenti parametri di input per evidenziare quali siano le differenze in termini di risultati. Per questo motivo sarà necessario, a seconda delle necessità, scegliere i corretti parametri temporali, di somiglianza e di generazione dei cluster.

In questo lavoro si è cercato di analizzare i dati con diversi parametri di input non solo per far emergere le differenze, ma anche per evidenziare le costanti. Elemento comune a tutte le analisi effettuate è la presenza di 4 clusters, ad elevata similarità intra-cluster, che, in accordo con lo standard GISC, è possibile chiamare: "Financials", "Utilities", "Energy Oil & Gas" e "Real Estate". Questi cluster contano in totale oltre 100 titoli sui 500 analizzati ed emergono da qualsiasi analisi anche con range temporali differenti.

Oltre al clustering si è reputato interessante dare spazio anche all'analisi della similarità tra singoli titoli. Da questa tipologia di analisi si è riscontrato, ad esempio, che i titoli più simili ad Apple (AAPL) siano aziende produttori di semiconduttori. Oppure come le aziende più vicine ad American Airlines (AAL) siano, dopo le altre compagnie aeree, quelle che gravitano attorno al mondo dei viaggi. Leggendo i risultati è possibile fare tantissimi esempi di questo tipo.

Vista la tecnica di definizione dei cluster a partire dalla matrice di similarità è stato anche possibile fare emergere i diversi livelli di relazione tra titoli di uno stesso GISC Sector. Ad esempio, fra tutti i titoli del settore "Helth Care", sono stati individuati diversi sub-cluster a cui è stato possibile dare un nome grazie alla definizione delle "GISC Industries" della tassonomia GISC. I risultati di questo tipo di analisi evidenziano anche outier o titoli molto simili che appartengono a Industries differenti. La visualizzazione tabellare di questi dati spesso non è di facile comprensione, per questo motivo è stata ideato e proposto un nuovo metodo di visualization che renda le analisi chiare e fruibili da chiunque.

Questo lavoro crea alcuni spunti interessanti per possibili sviluppi futuri. Nella prossima sezione si fa cenno di alcuni di questi.

5.2 Sviluppi futuri

5.2.1 Analisi globale dei mercati

In un mercato globale come quello moderno gli investimenti non si concentrano su un solo indice azionario o su indici di una sola nazione. È necessario quindi condurre le analisi utilizzando i dati dei mercati azionari di tutto il mondo. L'ampliamento dell'analisi crea principalmente due sfide:

- **Tecnica**: gli strumenti e la metogologia sviluppata in questo lavoro è scalabile?
- **Analitica**: i mercati di tutto il mondo hanno orari differenti, come gestire queste differenze?

Il primo quesito ci pone davanti una delle problematiche cardine del mining. L'ampliamento dei mercati da analizzare genererebbe una enorme crescita del numero di serie temporali. Questo potrebbe generare sia un problema di complessita (che, come precedentemente analizzato, scala linearmente con il numero di serie temporali da analizzare) sia un problema di efficacia (la metodologia per come è stata sviluppata è adatta ad un'analisi con un numero di serie molto più ampio?).

Il secondo quesito ci obbliga ad indagare sulle correlazioni che possono avere i titoli su diversi mercati. In particolare bisognerà verificare se sia possibile, come fatto in questo lavoro, considerare le variazioni percentuali giornaliere in quanto i differenti fuso orari potrebbero generare fluttuazioni in momenti diversi delle varie giornate. Se ad esempio un cambiamento nelle condizioni di mercato, ad esemio una notizia, genera delle reazioni nella Borsa di New York alle ore 15.00 (quindi in piena fase di contrattazione) le reazioni sulla Borsa di Milano o di Shanghai verranno registrate il giorno dopo in quanto nello stesso momento le due borse avranno già chiuso le contrattazioni (a Milano infatti saranno le 21.00 mentre a Shanghai le 04.00). In questo caso quindi i titoli degli indici di Milano e di Shanghai registreranno le reazioni alla notizia il giorno dopo rispetto ai titoli dell'indice di New York. La stessa cosa naturalmente potrebbe succedere al contrario. È necessario quindi trovare un metodo per comparare le serie che tenga conto di queste differenze.

5.2.2 Valute e criptovalute

Visti i buoni risultati ottenuti dall'applicazione di questa metodologia potrebbe essere interessante estendere l'ambito di ricerca anche ad altri tipi di dati sempre in ambito finanziario. Il clustering di serie finanziarie è un tema, come anticipato, già affrontato in ambito accademico. Ad esempio è stato citato l'articolo "*Clustering of financial time series*" [31] che affronta la tematica del clustering di serie temporali che descrivono l'andamento dei tassi di cambio dell'Euro rispetto ad altre valute internazionali. Si potrebbero quindi applicare le tecniche descritte all'interno di questo lavoro a dataset che descrivono l'andamento dei tassi di cambio delle valute.

Potrebbe essere anche un interessante sviluppo l'applicazione su dataset che descrivono l'andamento dei prezzi delle criptovalute. Le criptovalute però hanno due principali differenze rispetto al mercato azionario:

1. maggiore volatilità;
2. contrattazioni h24.

Queste due differenze ci obbligano a valutare come gestire il campionamento per la generazione delle serie temporali. Scelta che con il mercato azionario è stata ovviamente dettata dagli orari di contrattazione ma che nel caso delle criptovalute è da valutare.

Appendice A

Global Industry Classification Standard (GISC)

Nelle pagine successive la classificazione completa degli 11 sectors, 24 industry groups, 69 industries e 158 sub-industries della tassonomia GISC. La versione è l'ultima disponibile (settembre 2018).

Sector		Industry Group		Industry		Sub-Industry	
10	Energy	1010	Energy	101010	Energy Equipment & Services	10101010	Oil & Gas Drilling
						10101020	Oil & Gas Equipment & Services
				101020	Oil, Gas & Consumable Fuels	10102010	Integrated Oil & Gas
						10102020	Oil & Gas Exploration & Production
						10102030	Oil & Gas Refining & Marketing
						10102040	Oil & Gas Storage & Transportation
						10102050	Coal & Consumable Fuels
15	Materials	1510	Materials	151010	Chemicals	15101010	Commodity Chemicals
						15101020	Diversified Chemicals
						15101030	Fertilizers & Agricultural Chemicals
						15101040	Industrial Gases
						15101050	Specialty Chemicals
				151020	Construction Materials	15102010	Construction Materials
				151030	Containers & Packaging	15103010	Metal & Glass Containers
						15103020	Paper Packaging
				151040	Metals & Mining	15104010	Aluminum
						15104020	Diversified Metals & Mining
						15104025	Copper
						15104030	Gold
						15104040	Precious Metals & Minerals
						15104045	Silver
						15104050	Steel
				151050	Paper & Forest Products	15105010	Forest Products
						15105020	Paper Products
20	Industrials	2010	Capital Goods	201010	Aerospace & Defense	20101010	Aerospace & Defense
				201020	Building Products	20102010	Building Products
				201030	Construction & Engineering	20103010	Construction & Engineering
				201040	Electrical Equipment	20104010	Electrical Components & Equipment
						20104020	Heavy Electrical Equipment
				201050	Industrial Conglomerates	20105010	Industrial Conglomerates
				201060	Machinery	20106010	Construction Machinery & Heavy Trucks
						20106015	Agricultural & Farm Machinery
						20106020	Industrial Machinery
				201070	Trading Companies & Distributors	20107010	Trading Companies & Distributors
		2020	Commercial & Professional Services	202010	Commercial Services & Supplies	20201010	Commercial Printing
						20201050	Environmental & Facilities Services
						20201060	Office Services & Supplies
						20201070	Diversified Support Services
						20201080	Security & Alarm Services
				202020	Professional Services	20202010	Human Resource & Employment Services
						20202020	Research & Consulting Services
		2030	Transportation	203010	Air Freight & Logistics	20301010	Air Freight & Logistics
				203020	Airlines	20302010	Airlines
				203030	Marine	20303010	Marine
				203040	Road & Rail	20304010	Railroads
						20304020	Trucking
				203050	Transportation Infrastructure	20305010	Airport Services
						20305020	Highways & Railtracks
						20305030	Marine Ports & Services

25	Consumer Discretionary	2510	Automobiles & Components	251010	Auto Components	25101010	Auto Parts & Equipment
						25101020	Tires & Rubber
				251020	Automobiles	25102010	Automobile Manufacturers
						25102020	Motorcycle Manufacturers
		2520	Consumer Durables & Apparel	252010	Household Durables	25201010	Consumer Electronics
						25201020	Home Furnishings
						25201030	Homebuilding
						25201040	Household Appliances
						25201050	Housewares & Specialties
				252020	Leisure Products	25202010	Leisure Products
				252030	Textiles, Apparel & luxury goods	25203010	Apparel, Accessories & Luxury Goods
						25203020	Footwear
						25203030	Textiles
		2530	Consumer Services	253010	Hotels, Restaurants & Leisure	25301010	Casinos & Gaming
						25301020	Hotels, Resorts & Cruise Lines
						25301030	Leisure Facilities
						25301040	Restaurants
				253020	Diversified Consumer Services	25302010	Education Services
						25302020	Specialized Consumer Services
		2550	Retailing	255010	Distributors	25501010	Distributors
				255020	Internet & Direct Marketing Retail	25502020	Internet & Direct Marketing Retail
				255030	Multiline Retail	25503010	Department Stores
						25503020	General Merchandise Stores
				255040	Specialty Retail	25504010	Apparel Retail
						25504020	Computer & Electronics Retail
						25504030	Home Improvement Retail
						25504040	Specialty Stores
						25504050	Automotive Retail
						25504060	Homefurnishing Retail
30	Consumer Staples	3010	Food & Staples Retailing	301010	Food & Staples Retailing	30101010	Drug Retail
						30101020	Food Distributors
						30101030	Food Retail
						30101040	Hypermarkets & Super Centers
		3020	Food, Beverage & Tobacco	302010	Beverages	30201010	Brewers
						30201020	Distillers & Vintners
						30201030	Soft Drinks
				302020	Food Products	30202010	Agricultural Products
						30202030	Packaged Foods & Meats
				302030	Tobacco	30203010	Tobacco
		3030	Household & Personal Products	303010	Household Products	30301010	Household Products
				303020	Personal Products	30302010	Personal Products
35	Health Care	3510	Health Care Equipment & Services	351010	Health Care Equipment & Supplies	35101010	Health Care Equipment
						35101020	Health Care Supplies
				351020	Health Care Providers & Services	35102010	Health Care Distributors
						35102015	Health Care Services
						35102020	Health Care Facilities
						35102030	Managed Health Care
				351030	Health Care Technology	35103010	Health Care Technology
		3520	Pharmaceuticals, Biotechnology & Life Sciences	352010	Biotechnology	35201010	Biotechnology
				352020	Pharmaceuticals	35202010	Pharmaceuticals
				352030	Life Sciences Tools & Services	35203010	Life Sciences Tools & Services
40	Financials	4010	Banks	401010	Banks	40101010	Diversified Banks
						40101015	Regional Banks
				401020	Thrifts & Mortgage Finance	40102010	Thrifts & Mortgage Finance
		4020	Diversified Financials	402010	Diversified Financial Services	40201020	Other Diversified Financial Services
						40201030	Multi-Sector Holdings
						40201040	Specialized Finance
				402020	Consumer Finance	40202010	Consumer Finance

				402030	Capital Markets	40203010	Asset Management & Custody Banks
						40203020	Investment Banking & Brokerage
						40203030	Diversified Capital Markets
						40203040	Financial Exchanges & Data
				402040	Mortgage Real Estate Investment Trusts (REITs)	40204010	Mortgage REITs
		4030	Insurance	403010	Insurance	40301010	Insurance Brokers
						40301020	Life & Health Insurance
						40301030	Multi-line Insurance
						40301040	Property & Casualty Insurance
						40301050	Reinsurance
45	Information Technology	4510	Software & Services	451020	IT Services	45102010	IT Consulting & Other Services
						45102020	Data Processing & Outsourced Services
						45102030	Internet Services & Infrastructure
				451030	Software	45103010	Application Software
						45103020	Systems Software
		4520	Technology Hardware & Equipment	452010	Communications Equipment	45201020	Communications Equipment
				452020	Technology Hardware, Storage & Peripherals	45202030	Technology Hardware, Storage & Peripherals
				452030	Electronic Equipment, Instruments & Components	45203010	Electronic Equipment & Instruments
						45203015	Electronic Components
						45203020	Electronic Manufacturing Services
						45203030	Technology Distributors
		4530	Semiconductors & Semiconductor Equipment	453010	Semiconductors & Semiconductor Equipment	45301010	Semiconductor Equipment
						45301020	Semiconductors
50	Communication Services	5010	Telecommunication Services	501010	Diversified Telecommunication Services	50101010	Alternative Carriers
						50101020	Integrated Telecommunication Services
				501020	Wireless Telecommunication Services	50102010	Wireless Telecommunication Services
		5020	Media & Entertainment	502010	Media	50201010	Advertising
						50201020	Broadcasting
						50201030	Cable & Satellite
						50201040	Publishing
				502020	Entertainment	50202010	Movies & Entertainment
						50202020	Interactive Home Entertainment
				502030	Interactive Media & Services	50203010	Interactive Media & Services
55	Utilities	5510	Utilities	551010	Electric Utilities	55101010	Electric Utilities
				551020	Gas Utilities	55102010	Gas Utilities
				551030	Multi-Utilities	55103010	Multi-Utilities
				551040	Water Utilities	55104010	Water Utilities
				551050	Independent Power and Renewable Electricity Producers	55105010	Independent Power Producers & Energy Traders
						55105020	Renewable Electricity
60	Real Estate	6010	Real Estate	601010	Equity Real Estate Investment Trusts (REITs)	60101010	Diversified REITs
						60101020	Industrial REITs
						60101030	Hotel & Resort REITs
						60101040	Office REITs
						60101050	Health Care REITs
						60101060	Residential REITs
						60101070	Retail REITs
						60101080	Specialized REITs
				601020	Real Estate Management & Development	60102010	Diversified Real Estate Activities
						60102020	Real Estate Operating Companies
						60102030	Real Estate Development
						60102040	Real Estate Services

Appendice B

Aziende S&P500 con GISC Sub-Industry

Nelle pagine successive si riporta la lista completa di tutti i titoli presenti all'interno dello S&P500 e la rispettiva classificazione del "GISC Sub Industry" della tassonomia GISC. Questa tabella può essere di supporto nella lettura e nella comprensione dei risultati riportati all'interno di questo lavoro.

Simbolo	Nome	GICS Sub Industry
A	Agilent Technologies Inc	Health Care Equipment
AAL	American Airlines Group	Airlines
AAP	Advance Auto Parts	Automotive Retail
AAPL	Apple Inc.	Technology Hardware, Storage & Peripherals
ABBV	AbbVie Inc.	Pharmaceuticals
ABC	AmerisourceBergen Corp	Health Care Distributors
ABMD	ABIOMED Inc	Health Care Equipment
ABT	Abbott Laboratories	Health Care Equipment
ACN	Accenture plc	IT Consulting & Other Services
ADBE	Adobe Systems Inc	Application Software
ADI	Analog Devices, Inc.	Semiconductors
ADM	Archer-Daniels-Midland Co	Agricultural Products
ADP	Automatic Data Processing	Internet Services & Infrastructure
ADS	Alliance Data Systems	Data Processing & Outsourced Services
ADSK	Autodesk Inc.	Application Software
AEE	Ameren Corp	Multi-Utilities
AEP	American Electric Power	Electric Utilities
AES	AES Corp	Independent Power Producers & Energy Traders
AFL	AFLAC Inc	Life & Health Insurance
AGN	Allergan, Plc	Pharmaceuticals
AIG	American International Group	Property & Casualty Insurance
AIV	Apartment Investment & Management	Residential REITs
AIZ	Assurant	Multi-line Insurance
AJG	Arthur J. Gallagher & Co.	Insurance Brokers
AKAM	Akamai Technologies Inc	Internet Services & Infrastructure
ALB	Albemarle Corp	Specialty Chemicals
ALGN	Align Technology	Health Care Supplies
ALK	Alaska Air Group Inc	Airlines
ALL	Allstate Corp	Property & Casualty Insurance
ALLE	Allegion	Building Products
ALXN	Alexion Pharmaceuticals	Biotechnology
AMAT	Applied Materials Inc.	Semiconductor Equipment
AMD	Advanced Micro Devices Inc	Semiconductors
AME	AMETEK Inc.	Electrical Components & Equipment
AMG	Affiliated Managers Group Inc	Asset Management & Custody Banks
AMGN	Amgen Inc.	Biotechnology
AMP	Ameriprise Financial	Asset Management & Custody Banks
AMT	American Tower Corp.	Specialized REITs
AMZN	Amazon.com Inc.	Internet & Direct Marketing Retail
ANET	Arista Networks	Communications Equipment
ANSS	ANSYS	Application Software
ANTM	Anthem	Managed Health Care
AON	Aon plc	Insurance Brokers
AOS	A.O. Smith Corp	Building Products
APA	Apache Corporation	Oil & Gas Exploration & Production
APC	Anadarko Petroleum Corp	Oil & Gas Exploration & Production
APD	Air Products & Chemicals Inc	Industrial Gases
APH	Amphenol Corp	Electronic Components
APTV	Aptiv Plc	Auto Parts & Equipment
ARE	Alexandria Real Estate Equities	Office REITs
ARNC	Arconic Inc.	Aerospace & Defense
ATVI	Activision Blizzard	Interactive Home Entertainment
AVB	AvalonBay Communities, Inc.	Residential REITs
AVGO	Broadcom Inc.	Semiconductors
AVY	Avery Dennison Corp	Paper Packaging
AWK	American Water Works Company Inc	Water Utilities
AXP	American Express Co	Consumer Finance
AZO	AutoZone Inc	Specialty Stores
BA	Boeing Company	Aerospace & Defense
BAC	Bank of America Corp	Diversified Banks
BAX	Baxter International Inc.	Health Care Equipment
BBT	BB&T Corporation	Regional Banks
BBY	Best Buy Co. Inc.	Computer & Electronics Retail
BDX	Becton Dickinson	Health Care Equipment
BEN	Franklin Resources	Asset Management & Custody Banks
BF-B	Brown-Forman Corp.	Distillers & Vintners
BHGE	Baker Hughes, a GE Company	Oil & Gas Equipment & Services
BIIB	Biogen Inc.	Biotechnology
BK	The Bank of New York Mellon Corp.	Asset Management & Custody Banks
BKNG	Booking Holdings Inc	Internet & Direct Marketing Retail
BLK	BlackRock	Asset Management & Custody Banks
BLL	Ball Corp	Metal & Glass Containers
BMY	Bristol-Myers Squibb	Health Care Distributors
BR	Broadridge Financial Solutions	Data Processing & Outsourced Services
BRK-B	Berkshire Hathaway	Multi-Sector Holdings
BSX	Boston Scientific	Health Care Equipment
BWA	BorgWarner	Auto Parts & Equipment
BXP	Boston Properties	Office REITs
C	Citigroup Inc.	Diversified Banks
CAG	Conagra Brands	Packaged Foods & Meats
CAH	Cardinal Health Inc.	Health Care Distributors
CAT	Caterpillar Inc.	Construction Machinery & Heavy Trucks
CB	Chubb Limited	Property & Casualty Insurance
CBOE	Cboe Global Markets	Financial Exchanges & Data
CBRE	CBRE Group	Real Estate Services
CBS	CBS Corp.	Broadcasting
CCI	Crown Castle International Corp.	Specialized REITs
CCL	Carnival Corp.	Hotels, Resorts & Cruise Lines
CDNS	Cadence Design Systems	Application Software
CELG	Celgene Corp.	Biotechnology
CERN	Cerner	Health Care Technology
CF	CF Industries Holdings Inc	Fertilizers & Agricultural Chemicals
CFG	Citizens Financial Group	Regional Banks
CHD	Church & Dwight	Household Products
CHRW	C. H. Robinson Worldwide	Air Freight & Logistics
CHTR	Charter Communications	Cable & Satellite
CI	CIGNA Corp.	Managed Health Care
CINF	Cincinnati Financial	Property & Casualty Insurance
CL	Colgate-Palmolive	Household Products
CLX	The Clorox Company	Household Products
CMA	Comerica Inc.	Diversified Banks
CMCSA	Comcast Corp.	Cable & Satellite
CME	CME Group Inc.	Financial Exchanges & Data
CMG	Chipotle Mexican Grill	Restaurants

CMI	Cummins Inc.	Industrial Machinery
CMS	CMS Energy	Multi-Utilities
CNC	Centene Corporation	Managed Health Care
CNP	CenterPoint Energy	Multi-Utilities
COF	Capital One Financial	Consumer Finance
COG	Cabot Oil & Gas	Oil & Gas Exploration & Production
COO	The Cooper Companies	Health Care Supplies
COP	ConocoPhillips	Oil & Gas Exploration & Production
COST	Costco Wholesale Corp.	Hypermarkets & Super Centers
COTY	Coty, Inc	Personal Products
CPB	Campbell Soup	Packaged Foods & Meats
CPRT	Copart Inc	Diversified Support Services
CRM	Salesforce.com	Internet Software & Services
CSCO	Cisco Systems	Communications Equipment
CSX	CSX Corp.	Railroads
CTAS	Cintas Corporation	Diversified Support Services
CTL	CenturyLink Inc	Integrated Telecommunication Services
CTSH	Cognizant Technology Solutions	IT Consulting & Other Services
CTXS	Citrix Systems	Internet Software & Services
CVS	CVS Health	Health Care Services
CVX	Chevron Corp.	Integrated Oil & Gas
CXO	Concho Resources	Oil & Gas Exploration & Production
D	Dominion Energy	Electric Utilities
DAL	Delta Air Lines Inc.	Airlines
DE	Deere & Co.	Agricultural & Farm Machinery
DFS	Discover Financial Services	Consumer Finance
DG	Dollar General	General Merchandise Stores
DHI	D. R. Horton	Homebuilding
DHR	Danaher Corp.	Health Care Equipment
DIS	The Walt Disney Company	Movies & Entertainment
DISCA	Discovery Inc. Class A	Broadcasting
DISCK	Discovery Inc. Class C	Broadcasting
DISH	Dish Network	Cable & Satellite
DLR	Digital Realty Trust Inc	Specialized REITs
DLTR	Dollar Tree	General Merchandise Stores
DOV	Dover Corp.	Industrial Machinery
DRE	Duke Realty Corp	Industrial REITs
DRI	Darden Restaurants	Restaurants
DTE	DTE Energy Co.	Multi-Utilities
DUK	Duke Energy	Electric Utilities
DVA	DaVita Inc.	Health Care Facilities
DVN	Devon Energy	Oil & Gas Exploration & Production
DWDP	Rimossa	Rimossa
DXC	DXC Technology	IT Consulting & Other Services
EA	Electronic Arts	Interactive Home Entertainment
EBAY	eBay Inc.	Internet & Direct Marketing Retail
ECL	Ecolab Inc.	Specialty Chemicals
ED	Consolidated Edison	Electric Utilities
EFX	Equifax Inc.	Research & Consulting Services
EIX	Edison Int'l	Electric Utilities
EL	Estee Lauder Cos.	Personal Products
EMN	Eastman Chemical	Diversified Chemicals
EMR	Emerson Electric Company	Electrical Components & Equipment
EOG	EOG Resources	Oil & Gas Exploration & Production
EQIX	Equinix	Specialized REITs
EQR	Equity Residential	Residential REITs
ES	Eversource Energy	Multi-Utilities
ESRX	Rimossa	Rimossa
ESS	Essex Property Trust, Inc.	Residential REITs
ETFC	E*Trade	Investment Banking & Brokerage
ETN	Eaton Corporation	Electrical Components & Equipment
ETR	Entergy Corp.	Electric Utilities
EW	Edwards Lifesciences	Health Care Equipment
EXC	Exelon Corp.	Multi-Utilities
EXPD	Expeditors	Air Freight & Logistics
EXPE	Expedia Group	Internet & Direct Marketing Retail
EXR	Extra Space Storage	Specialized REITs
F	Ford Motor	Automobile Manufacturers
FANG	Diamondback Energy	Oil & Gas Exploration & Production
FAST	Fastenal Co	Building Products
FB	Facebook, Inc.	Interactive Media & Services
FBHS	Fortune Brands Home & Security	Building Products
FCX	Freeport-McMoRan Inc.	Copper
FDX	FedEx Corporation	Air Freight & Logistics
FE	FirstEnergy Corp	Electric Utilities
FFIV	F5 Networks	Communications Equipment
FIS	Fidelity National Information Services	Internet Software & Services
FISV	Fiserv Inc	Internet Software & Services
FITB	Fifth Third Bancorp	Regional Banks
FL	Foot Locker Inc	Apparel Retail
FLIR	FLIR Systems	Electronic Equipment & Instruments
FLR	Rimossa	Rimossa
FLS	Flowserve Corporation	Industrial Machinery
FLT	FleetCor Technologies Inc	Data Processing & Outsources Services
FMC	FMC Corporation	Fertilizers & Agricultural Chemicals
FOX	Fox Corporation Class B	Movies & Entertainment
FOXA	Fox Corporation Class A	Movies & Entertainment
FRT	Federal Realty Investment Trust	Retail REITs
FTI	TechnipFMC	Oil & Gas Equipment & Services
FTNT	Fortinet	Systems Software
GD	General Dynamics	Aerospace & Defense
GE	General Electric	Industrial Conglomerates
GILD	Gilead Sciences	Biotechnology
GIS	General Mills	Packaged Foods & Meats
GLW	Corning Inc.	Electronic Components
GM	General Motors	Automobile Manufacturers
GOOG	Alphabet Inc Class C	Interactive Media & Services
GOOGL	Alphabet Inc Class A	Interactive Media & Services
GPC	Genuine Parts	Specialty Stores
GPN	Global Payments Inc.	Data Processing & Outsourced Services
GPS	Gap Inc.	Apparel Retail
GRMN	Garmin Ltd.	Consumer Electronics
GS	Goldman Sachs Group	Investment Banking & Brokerage
GT	Rimossa	Rimossa
GWW	Grainger (W.W.) Inc.	Industrial Machinery
HAL	Halliburton Co.	Oil & Gas Equipment & Services
HAS	Hasbro Inc.	Leisure Products
HBAN	Huntington Bancshares	Regional Banks
HBI	Hanesbrands Inc	Apparel, Accessories & Luxury Goods

HCA	HCA Healthcare	Health Care Facilities
HCP	HCP Inc.	Health Care REITs
HD	Home Depot	Home Improvement Retail
HES	Hess Corporation	Integrated Oil & Gas
HFC	HollyFrontier Corp	Oil & Gas Refining & Marketing
HIG	Hartford Financial Svc.Gp.	Property & Casualty Insurance
HII	Huntington Ingalls Industries	Aerospace & Defense
HLT	Hilton Worldwide Holdings Inc	Hotels, Resorts & Cruise Lines
HOG	Harley-Davidson	Motorcycle Manufacturers
HOLX	Hologic	Health Care Equipment
HON	Honeywell Int'l Inc.	Industrial Conglomerates
HP	Helmerich & Payne	Oil & Gas Drilling
HPE	Hewlett Packard Enterprise	Technology Hardware, Storage & Peripherals
HPQ	HP Inc.	Technology Hardware, Storage & Peripherals
HRB	Block H&R	Specialized Consumer Services
HRL	Hormel Foods Corp.	Packaged Foods & Meats
HRS	#N/D	#N/D
HSIC	Henry Schein	Health Care Distributors
HST	Host Hotels & Resorts	Hotel & Resort REITs
HSY	The Hershey Company	Packaged Foods & Meats
HUM	Humana Inc.	Managed Health Care
IBM	International Business Machines	IT Consulting & Other Services
ICE	Intercontinental Exchange	Financial Exchanges & Data
IDXX	IDEXX Laboratories	Health Care Equipment
IFF	Intl Flavors & Fragrances	Specialty Chemicals
ILMN	Illumina Inc	Life Sciences Tools & Services
INCY	Incyte	Biotechnology
INFO	IHS Markit Ltd.	Research & Consulting Services
INTC	Intel Corp.	Semiconductors
INTU	Intuit Inc.	Internet Software & Services
IP	International Paper	Paper Packaging
IPG	Interpublic Group	Advertising
IPGP	IPG Photonics Corp.	Electronic Manufacturing Services
IQV	IQVIA Holdings Inc.	Life Sciences Tools & Services
IR	Ingersoll-Rand PLC	Industrial Machinery
IRM	Iron Mountain Incorporated	Specialized REITs
ISRG	Intuitive Surgical Inc.	Health Care Equipment
IT	Gartner Inc	IT Consulting & Other Services
ITW	Illinois Tool Works	Industrial Machinery
IVZ	Invesco Ltd.	Asset Management & Custody Banks
JBHT	J. B. Hunt Transport Services	Trucking
JCI	Johnson Controls International	Building Products
JEC	Jacobs Engineering Group	Construction & Engineering
JKHY	Jack Henry & Associates	Data Processing & Outsourced Services
JNJ	Johnson & Johnson	Pharmaceuticals
JNPR	Juniper Networks	Communications Equipment
JPM	JPMorgan Chase & Co.	Diversified Banks
JWN	Nordstrom	Department Stores
K	Kellogg Co.	Packaged Foods & Meats
KEY	KeyCorp	Regional Banks
KEYS	Keysight Technologies	Electronic Equipment & Instruments
KHC	Kraft Heinz Co	Packaged Foods & Meats
KIM	Kimco Realty	Retail REITs
KLAC	KLA Corporation	Semiconductor Equipment
KMB	Kimberly-Clark	Household Products
KMI	Kinder Morgan	Oil & Gas Storage & Transportation
KMX	Carmax Inc	Specialty Stores
KO	Coca-Cola Company	Soft Drinks
KORS	Michael Kors	Apparel, Accessories & Luxury Goods
KR	Kroger Co.	Food Retail
KSS	Kohl's Corp.	General Merchandise Stores
KSU	Kansas City Southern	Railroads
L	Loews Corp.	Multi-line Insurance
LB	L Brands Inc.	Apparel Retail
LEG	Leggett & Platt	Home Furnishings
LEN	Lennar Corp.	Homebuilding
LH	Laboratory Corp. of America Holding	Health Care Services
LKQ	LKQ Corporation	Distributors
LLL	#N/D	#N/D
LLY	Lilly (Eli) & Co.	Pharmaceuticals
LMT	Lockheed Martin Corp.	Aerospace & Defense
LNC	Lincoln National	Multi-line Insurance
LNT	Alliant Energy Corp	Electric Utilities
LOW	Lowe's Cos.	Home Improvement Retail
LRCX	Lam Research	Semiconductor Equipment
LUV	Southwest Airlines	Airlines
LYB	LyondellBasell	Specialty Chemicals
M	Macy's Inc.	Department Stores
MA	Mastercard Inc.	IT Services
MAA	Mid-America Apartments	Residential REITs
MAC	Macerich	Retail REITs
MAR	Marriott Int'l.	Hotels, Resorts & Cruise Lines
MAS	Masco Corp.	Building Products
MAT	Rimossa	Rimossa
MCD	McDonald's Corp.	Restaurants
MCHP	Microchip Technology	Semiconductors
MCK	McKesson Corp.	Health Care Distributors
MCO	Moody's Corp	Financial Exchanges & Data
MDLZ	Mondelez International	Packaged Foods & Meats
MDT	Medtronic plc	Health Care Equipment
MET	MetLife Inc.	Life & Health Insurance
MGM	MGM Resorts International	Casinos & Gaming
MHK	Mohawk Industries	Home Furnishings
MKC	McCormick & Co.	Packaged Foods & Meats
MLM	Martin Marietta Materials	Construction Materials
MMC	Marsh & McLennan	Insurance Brokers
MMM	3M Company	Industrial Conglomerates
MNST	Monster Beverage	Soft Drinks
MO	Altria Group Inc	Tobacco
MOS	The Mosaic Company	Fertilizers & Agricultural Chemicals
MPC	Marathon Petroleum	Oil & Gas Refining & Marketing
MRK	Merck & Co.	Pharmaceuticals
MRO	Marathon Oil Corp.	Oil & Gas Exploration & Production
MS	Morgan Stanley	Investment Banking & Brokerage
MSCI	MSCI Inc	Financial Exchanges & Data
MSFT	Microsoft Corp.	Systems Software
MSI	Motorola Solutions Inc.	Communications Equipment
MTB	M&T Bank Corp.	Regional Banks
MTD	Mettler Toledo	Life Sciences Tools & Services

MU	Micron Technology	Semiconductors
MXIM	Maxim Integrated Products Inc	Semiconductors
MYL	Mylan N.V.	Pharmaceuticals
nan	#N/D	#N/D
NBL	Noble Energy Inc	Oil & Gas Exploration & Production
NCLH	Norwegian Cruise Line Holdings	Hotels, Resorts & Cruise Lines
NDAQ	Nasdaq, Inc.	Financial Exchanges & Data
NEE	NextEra Energy	Multi-Utilities
NEM	Newmont Goldcorp	Gold
NFLX	Netflix Inc.	Movies & Entertainment
NFX	Rimossa	Rimossa
NI	NiSource Inc.	Multi-Utilities
NKE	Nike	Apparel, Accessories & Luxury Goods
NKTR	Nektar Therapeutics	Pharmaceuticals
NLSN	Nielsen Holdings	Research & Consulting Services
NOC	Northrop Grumman	Aerospace & Defense
NOV	National Oilwell Varco Inc.	Oil & Gas Equipment & Services
NRG	NRG Energy	Independent Power Producers & Energy Traders
NSC	Norfolk Southern Corp.	Railroads
NTAP	NetApp	Internet Software & Services
NTRS	Northern Trust Corp.	Asset Management & Custody Banks
NUE	Nucor Corp.	Steel
NVDA	Nvidia Corporation	Semiconductors
NWL	Newell Brands	Housewares & Specialties
NWS	News Corp. Class B	Publishing
NWSA	News Corp. Class A	Publishing
O	Realty Income Corporation	Retail REITs
OKE	ONEOK	Oil & Gas Storage & Transportation
OMC	Omnicom Group	Advertising
ORCL	Oracle Corp.	Application Software
ORLY	O'Reilly Automotive	Specialty Stores
OXY	Occidental Petroleum	Oil & Gas Exploration & Production
PAYX	Paychex Inc.	Internet Software & Services
PBCT	People's United Financial	Thrifts & Mortgage Finance
PCAR	PACCAR Inc.	Construction Machinery & Heavy Trucks
PCG	Rimossa	Rimossa
PEG	Public Serv. Enterprise Inc.	Electric Utilities
PEP	PepsiCo Inc.	Soft Drinks
PFE	Pfizer Inc.	Pharmaceuticals
PFG	Principal Financial Group	Life & Health Insurance
PG	Procter & Gamble	Personal Products
PGR	Progressive Corp.	Property & Casualty Insurance
PH	Parker-Hannifin	Industrial Machinery
PHM	Pulte Homes Inc.	Homebuilding
PKG	Packaging Corporation of America	Paper Packaging
PKI	PerkinElmer	Health Care Equipment
PLD	Prologis	Industrial REITs
PM	Philip Morris International	Tobacco
PNC	PNC Financial Services	Regional Banks
PNR	Pentair plc	Industrial Machinery
PNW	Pinnacle West Capital	Multi-Utilities
PPL	PPL Corp.	Electric Utilities
PRGO	Perrigo	Pharmaceuticals
PRU	Prudential Financial	Life & Health Insurance
PSA	Public Storage	Specialized REITs
PSX	Phillips 66	Oil & Gas Refining & Marketing
PVH	PVH Corp.	Apparel, Accessories & Luxury Goods
PWR	Quanta Services Inc.	Construction & Engineering
PXD	Pioneer Natural Resources	Oil & Gas Exploration & Production
PYPL	PayPal	Data Processing & Outsourced Services
QCOM	QUALCOMM Inc.	Semiconductors
QRVO	Qorvo	Semiconductors
RCL	Royal Caribbean Cruises Ltd	Hotels, Resorts & Cruise Lines
RE	Everest Re Group Ltd.	Reinsurance
REG	Regency Centers Corporation	Retail REITs
REGN	Regeneron Pharmaceuticals	Biotechnology
RF	Regions Financial Corp.	Regional Banks
RHI	Robert Half International	Human Resource & Employment Services
RHT	#N/D	#N/D
RJF	Raymond James Financial Inc.	Investment Banking & Brokerage
RL	Ralph Lauren Corporation	Apparel, Accessories & Luxury Goods
RMD	ResMed	Health Care Equipment
ROK	Rockwell Automation Inc.	Electrical Components & Equipment
ROL	Rollins Inc.	Environmental & Facilities Services
ROP	Roper Technologies	Industrial Conglomerates
ROST	Ross Stores	Apparel Retail
RSG	Republic Services Inc	Environmental & Facilities Services
RTN	Raytheon Co.	Aerospace & Defense
SBAC	SBA Communications	Specialized REITs
SBUX	Starbucks Corp.	Restaurants
SCG	Rimossa	Rimossa
SCHW	Charles Schwab Corporation	Investment Banking & Brokerage
SEE	Sealed Air	Paper Packaging
SHW	Sherwin-Williams	Specialty Chemicals
SIVB	SVB Financial	Regional Banks
SJM	JM Smucker	Packaged Foods & Meats
SLB	Schlumberger Ltd.	Oil & Gas Equipment & Services
SLG	SL Green Realty	Office REITs
SNA	Snap-on	Industrial Machinery
SNPS	Synopsys Inc.	Application Software
SO	Southern Co.	Electric Utilities
SPG	Simon Property Group Inc	Retail REITs
SPGI	S&P Global, Inc.	Financial Exchanges & Data
SRE	Sempra Energy	Multi-Utilities
STI	SunTrust Banks	Regional Banks
STT	State Street Corp.	Asset Management & Custody Banks
STX	Seagate Technology	Technology Hardware, Storage & Peripherals
STZ	Constellation Brands	Distillers & Vintners
SWK	Stanley Black & Decker	Industrial Machinery
SWKS	Skyworks Solutions	Semiconductors
SYF	Synchrony Financial	Consumer Finance
SYK	Stryker Corp.	Health Care Equipment
SYMC	Symantec Corp.	Application Software
SYY	Sysco Corp.	Food Distributors

T	AT&T Inc.	Integrated Telecommunication Services
TAP	Molson Coors Brewing Company	Brewers
TDG	TransDigm Group	Aerospace & Defense
TEL	TE Connectivity Ltd.	Electronic Manufacturing Services
TGT	Target Corp.	General Merchandise Stores
TIF	Tiffany & Co.	Apparel, Accessories & Luxury Goods
TJX	TJX Companies Inc.	Apparel Retail
TMK	Torchmark Corp.	Life & Health Insurance
TMO	Thermo Fisher Scientific	Health Care Equipment
TPR	Tapestry, Inc.	Apparel, Accessories & Luxury Goods
TRIP	TripAdvisor	Interactive Media & Services
TROW	T. Rowe Price Group	Asset Management & Custody Banks
TRV	The Travelers Companies Inc.	Property & Casualty Insurance
TSCO	Tractor Supply Company	Specialty Stores
TSN	Tyson Foods	Packaged Foods & Meats
TSS	Total System Services	Internet Software & Services
TTWO	Take-Two Interactive	Interactive Home Entertainment
TWTR	Twitter, Inc.	Interactive Media & Services
TXN	Texas Instruments	Semiconductors
TXT	Textron Inc.	Aerospace & Defense
UA	Under Armour Class C	Apparel, Accessories & Luxury Goods
UAA	Under Armour Class A	Apparel, Accessories & Luxury Goods
UAL	United Airlines Holdings	Airlines
UDR	UDR, Inc.	Residential REITs
UHS	Universal Health Services, Inc.	Health Care Facilities
ULTA	Ulta Beauty	Specialty Stores
UNH	United Health Group Inc.	Managed Health Care
UNM	Unum Group	Life & Health Insurance
UNP	Union Pacific Corp	Railroads
UPS	United Parcel Service	Air Freight & Logistics
URI	United Rentals, Inc.	Trading Companies & Distributors
USB	U.S. Bancorp	Diversified Banks
UTX	United Technologies	Aerospace & Defense
V	Visa Inc.	IT Services
VAR	Varian Medical Systems	Health Care Equipment
VFC	V.F. Corp.	Apparel, Accessories & Luxury Goods
VIAB	Viacom Inc.	Movies & Entertainment
VLO	Valero Energy	Oil & Gas Refining & Marketing
VMC	Vulcan Materials	Construction Materials
VNO	Vornado Realty Trust	Office REITs
VRSK	Verisk Analytics	Research & Consulting Services
VRSN	Verisign Inc.	Internet Services & Infrastructure
VRTX	Vertex Pharmaceuticals Inc	Biotechnology
VTR	Ventas Inc	Health Care REITs
VZ	Verizon Communications	Integrated Telecommunication Services
WAT	Waters Corporation	Health Care Distributors
WBA	Walgreens Boots Alliance	Drug Retail
WCG	WellCare	Managed Health Care
WDC	Western Digital	Technology Hardware, Storage & Peripherals
WEC	Wec Energy Group Inc	Electric Utilities
WELL	Welltower Inc.	Health Care REITs
WFC	Wells Fargo	Diversified Banks
WHR	Whirlpool Corp.	Household Appliances
WLTW	Willis Towers Watson	Insurance Brokers
WM	Waste Management Inc.	Environmental & Facilities Services
WMB	Williams Cos.	Oil & Gas Storage & Transportation
WMT	Walmart	Hypermarkets & Super Centers
WRK	WestRock	Paper Packaging
WU	Western Union Co	Internet Software & Services
WY	Weyerhaeuser	Specialized REITS
WYNN	Wynn Resorts Ltd	Casinos & Gaming
XEC	Cimarex Energy	Oil & Gas Exploration & Production
XEL	Xcel Energy Inc	Multi-Utilities
XLNX	Xilinx	Semiconductors
XOM	Exxon Mobil Corp.	Integrated Oil & Gas
XRAY	Dentsply Sirona	Health Care Supplies
XRX	Xerox	Technology Hardware, Storage & Peripherals
XYL	Xylem Inc.	Industrial Machinery
YUM	Yum! Brands Inc	Restaurants
ZBH	Zimmer Biomet Holdings	Health Care Equipment
ZION	Zions Bancorp	Regional Banks
ZTS	Zoetis	Pharmaceuticals

Elenco delle figure

Elenco delle tabelle

Bibliografia

[1] J. Paparrizos and L. Gravano, "k-shape: Efficient and accurate clustering of time series", ACM SIGMOD Record, vol. 45, 06 2016, pp. 69–76, DOI 10.1145/2949741.2949758

[2] S. Aghabozorgi, A. S. Shirkhorshidi, and T. Wah, "Time-series clustering - a decade review", Information Systems, vol. 53, 05 2015, DOI 10.1016/j.is.2015.04.007

[3] U. Rebbapragada, P. Protopapas, C. Brodley, and C. Alcock, "Finding anomalous periodic time series: An application to catalogs of periodic variable stars", Machine Learning, vol. 74, 05 2009, DOI 10.1007/s10994-008-5093-3

[4] N. Subhani, L. Rueda, A. Ngom, and C. Burden, "Multiple gene expression profile alignment for microarray time-series data clustering", Bioinformatics (Oxford, England), vol. 26, 09 2010, pp. 2281–8, DOI 10.1093/bioinformatics/btq422

[5] A. Elangasinghe, N. Singhal, K. Dirks, J. Salmond, and S. Samarasinghe, "Complex time series analysis of pm10 and pm2.5 for a coastal site using artificial neural network modelling and k-means clustering", Atmospheric Environment, vol. 94, 09 2014, DOI 10.1016/j.atmosenv.2014.04.051

[6] F. Iglesias Vázquez and W. Kastner, "Analysis of similarity measures in times series clustering for the discovery of building energy patterns", Energies, vol. 6, 02 2013, pp. 579–597, DOI 10.3390/en6020579

[7] H.-S. Guan and Q. Jiang, "Cluster financial time series for portfolio", 12 2007, pp. 851 – 856, DOI 10.1109/ICWAPR.2007.4420788

[8] C. Guo, H. Jia, and N. Zhang, "Time series clustering based on ica for stock data analysis", 11 2008, pp. 1 – 4, DOI 10.1109/WiCom.2008.2534

[9] S. Aghabozorgi and T. Wah, "Stock market co-movement assessment using a three-phase clustering method", Expert Systems with Applications: An International Journal, vol. 41, 03 2014, pp. 1301–1314, DOI 10.1016/j.eswa.2013.08.028

[10] F. Gullo, G. Ponti, A. Tagarelli, G. Tradigo, and P. Veltri, "A time series approach for clustering mass spectrometry data", Journal of Computational Science, vol. 3, 09 2012, DOI 10.1016/j.jocs.2011.06.008

[11] V. Kurbalija, C. Von Bernstorff, H.-D. Burkhard, J. Nachtwei, M. Ivanovic, and L. Fodor, "Time-series mining in a psychological domain", 09 2012, pp. 58–63, DOI 10.1145/2371316.2371328

[12] M. Ramoni, P. Sebastiani, and P. Cohen, "Multivariate clustering by dynamics.", 01 2000, pp. 633–638

[13] D. Tran and M. Wagner, "Fuzzy c-means clustering-based speaker verification", 02 2002, pp. 318–324, DOI 10.1007/3-540-45631-7_42

[14] J. Zhu, B. Wang, and B. Wu, "Social network users clustering based on multivariate time series of emotional behavior", The Journal of China Universities of Posts and Telecommunications, vol. 21, 04 2014, p. 21?31, DOI 10.1016/S1005-8885(14)60282-X

[15] H. Ding, G. Trajcevski, P. Scheuermann, X. Wang, and E. Keogh, "Querying and mining of time series data: Experimental comparison of representations and distance measures", PVLDB, vol. 1, 08 2008, pp. 1542–1552

[16] R. Agrawal, C. Faloutsos, and A. Swami, "Efficient similarity search in sequence databases", 01 1993, pp. 69–84

[17] N. Basalto, R. Bellotti, F. Carlo, P. Facchi, and S. Pascazio, "Hausdorff clustering of financial time series", arXiv.org, Quantitative Finance Papers, vol. 379, 04 2005, DOI 10.1016/j.physa.2007.01.011

[18] F. Shao, S. Cai, and J. Gu, "A modified hausdorff distance based algorithm for 2-dimensional spatial trajectory matching", 09 2010, pp. 166 – 172, DOI 10.1109/ICCSE.2010.5593666
[19] P. Smyth, "Clustering sequences with hidden markov models", Advances in Neural Information Processing Systems, vol. 9, 07 1999
[20] A. Banerjee and J. Ghosh, "Clickstream clustering using weighted longest common subsequences", 08 2001
[21] U. Mori, A. Mendiburu, and J. Lozano, "Similarity measure selection for clustering time series databases", IEEE Transactions on Knowledge and Data Engineering, vol. 28, 01 2015, pp. 1–1, DOI 10.1109/TKDE.2015.2462369
[22] C. Cassisi, P. Montalto, M. Aliotta, A. Cannata, and A. Pulvirenti, "Similarity measures and dimensionality reduction techniques for time series data mining". 09 2012
[23] E. Keogh and S. Kasetty, "On the need for time series data mining benchmarks", 01 2002, p. 102, DOI 10.1145/775060.775062
[24] T. Gorecki, "Classification of time series using combination of dtw and lcss dissimilarity measures", Communications in Statistics - Simulation and Computation, vol. 47, 01 2017, DOI 10.1080/03610918.2017.1280829
[25] J. MacQueen, "Some methods for classification and analysis of multivariate observations", 01 1967, pp. 281–297
[26] L. Kaufman and P. Rousseeuw, "Finding groups in data: An introduction to cluster analysis", 01 1990, DOI 10.2307/2290430
[27] J. Dunn, "A fuzzy relative of the isodata process and its use in detecting compact well-separated clusters", Cybernetics and Systems, vol. 3, 11 1973, pp. 32–57, DOI 10.1080/01969727308546046
[28] R. Krishnapuram, A. Joshi, O. Nasraoui, and L. Yi, "Low-complexity fuzzy relational clustering algorithms for web mining", Fuzzy Systems, IEEE Transactions on, vol. 9, 09 2001, pp. 595 – 607, DOI 10.1109/91.940971
[29] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise", 01 1996, pp. 226–231
[30] J. Murphy, "Technical analysis of financial markets", 01 1999
[31] P. D'Urso, C. Cappelli, D. Lallo, and R. Massari, "Clustering of financial time series", Physica A: Statistical Mechanics and its Applications, vol. 392, 05 2013, p. 2114?2129, DOI 10.1016/j.physa.2013.01.027
[32] S. Focardi, "Clustering economic and financial time series: Exploring the existence of stable correlation conditions", 06 2002
[33] G. Marti, "Clustering financial time series: How long is enough?", 03 2016
[34] N. Patel and J. Woo, "Clustering seasonality patterns in the presence of errors", 08 2002, DOI 10.1145/775047.775129
[35] H. Markowitz, "Portfolio selection: Efficient diversification of investment", The Journal of Finance, vol. 15, 12 1959, DOI 10.2307/2326204
[36] F. Petitjean and P. Gancarski, "Summarizing a set of time series by averaging: From steiner sequence to compact multiple alignment", Theoretical Computer Science, vol. 414, 01 2012, pp. 76–, DOI 10.1016/j.tcs.2011.09.029
[37] J. Thalheim, "Python implementation of k-shape." https://github.com/johnpaparrizos/kshape, Sep 2017, Accessed on 2019-08-01
[38] "Yahoo finance." https://it.finance.yahoo.com/, Feb 2019
[39] CRSP, "Crsp data definitions." http://www.crsp.com/products/documentation/data-definitions-2, 2019
[40] The Global Industry Classification Standard (GICS), https://www.msci.com/gics