UNIVERSITA' DEGLI STUDI DELL'INSUBRIA

DIPARTIMENTO DI ECONOMIA

CORSO DI LAUREA IN ECONOMIA E MANAGEMENT

COME COMINCIA L'IMPRESA?

Gruppi di imprese e benchmark strategici: applicazione degli strumenti di clustering nelle analisi di settore

Relatore:
Valerio Langè

Candidato:
Federico Rausa,
mat.744473

A. A. 2022/23

Ai miei genitori
per la fiducia
Agli amici
per la presenza
A mia nonna
per il consiglio

Indice

Abstract

Le analisi di settore consentono alle imprese di effettuare scelte di gestione rilevanti, ma i metodi per svolgerle sono numerosi e non sempre includono l'adozione di strumenti di analisi dati.
Il presente lavoro intende identificare i pattern decisionali, basati su fatti esogeni, che consentono di massimizzare l'efficienza delle decisioni di investimento, acquisizione e localizzazione d'azienda, a partire dai bilanci e dalle informazioni delle imprese già presenti sul mercato e raccolte in statistiche il più possibile esaustive, precise e sintetiche.
Questo, attraverso algoritmi di clustering e funzioni di ranking che rendono possibile effettuare riduzioni dimensionali dei dataset del settore analizzato.
Il presente lavoro mostra che i cluster segmentano le diverse fasce di performance nella popolazione e consentono di identificarne i benchmark e i relativi pattern decisionali.
Il clustering, quindi, è uno strumento utile per osservare sinteticamente grandi moli di dati inerenti a uno specifico settore, e la sua elasticità lo rende idoneo a molteplici applicazioni in campo economico e gestionale.

1. Introduzione

Nel presente lavoro, si intende offrire uno strumento statistico per esplorare rapidamente le opzioni strategiche connesse a un determinato settore industriale.
In particolare, si assume la prospettiva di chi intende entrare in un nuovo mercato, avviando un'attività commerciale ex novo o mediante l'acquisizione di un'impresa già avviata.
Lo strumento tipico adottato per assolvere tale problema, almeno dal punto di vista dei fattori esogeni, è l'analisi di settore, che costituisce un tema vastamente affrontato dagli economisti.
In diverse materie economiche esiste una ricca letteratura di metodi per effettuare tale analisi, ma in pochi casi si ha un approccio orientato all'adozione di semplici strumenti statistici.
A partire dai dati di bilancio di centinaia o migliaia di imprese, operanti nel medesimo settore o nella medesima filiera produttiva, risulta complesso ricavare una visione sintetica delle strategie comunemente più adottate e valutare la bontà dei relativi esiti.
Diversi autori[1] propongono l'applicazione di algoritmi di clustering alle analisi di settore, per classificare in modo efficiente le imprese accomunate da una certa caratteristica, come l'appartenenza al medesimo mercato, alla medesima area geografica o alla medesima supply chain.
Un algoritmo di clustering, ovvero di raggruppamento, consente di suddividere una popolazione di individui, o osservazioni, in un certo numero di cluster, ovvero di sottogruppi, in base al grado di somiglianza quantitativa tra ogni coppia di osservazioni[2].
Si tratta di un insieme di strumenti che hanno trovato numerose applicazioni negli ultimi anni: vengono utilizzati nei motori di ricerca[3] , nei sistemi di raccomandazione di contenuti per gli utenti (come gli ads pubblicitari di Google, i social network e altre piattaforme digitali come Amazon e Youtube), nella ricerca biologica e medica, nelle analisi di suoni e immagini e nella segmentazione dei consumatori in operazioni di marketing[4].
Nelle analisi dei bilanci di più imprese, i modelli di clustering consentono di separare le imprese di piccole, medie e grandi dimensioni, identificare frodi fiscali[5], costruire dei valori di benchmark e segmentare le diverse fasce di mercato per strategia operativa o natura dell'attività[6].

---

[1] Benabdellah, A. C., Benghabrit, A., & Bouhaddou, I. (2019). A survey of clustering algorithms for an industrial context. *Procedia computer science*, *148*, 291-302.
Song, L., Dong, Y., Guo, Q., Meng, Y., & Zhao, G. (2023). An adaptive differential evolution algorithm with DBSCAN for the integrated slab allocation problem in steel industry. *Applied Soft Computing*, *146*, 110665.
Liu, G., Yang, J., Hao, Y., & Zhang, Y. (2018). Big data-informed energy efficiency assessment of China industry sectors based on K-means clustering. *Journal of cleaner production*, *183*, 304-314.
[2] Oyewole, G. J., & Thopil, G. A. (2023). Data clustering: Application and trends. *Artificial Intelligence Review*, *56*(7), 6439-6475.
[3] Mecca, G., Raunich, S., & Pappalardo, A. (2007). A new algorithm for clustering search results. *Data & Knowledge Engineering*, *62*(3), 504-522.
[4] Kuo, R. J., Ho, L. M., & Hu, C. M. (2002). Integration of self-organizing feature map and K-means algorithm for market segmentation. *Computers & Operations Research*, *29*(11), 1475-1493.
[5] Deng, Q., & Mei, G. (2009, August). Combining self-organizing map and k-means clustering for detecting fraudulent financial statements. In *2009 IEEE international conference on granular computing* (pp. 126-131). IEEE.
Sabau, A. S. (2012). Survey of clustering based financial fraud detection research. *Informatica Economica*, *16*(1), 110.
[6] Chong, D., & Zhu, H. H. (2012, December). Firm clustering based on financial statements. In *22nd Workshop on Information Technology and Information Systems (WITS'12)*.

Il clustering si presenta quindi come un ausilio statistico molto utile per effettuare analisi di settore, in grado di integrare e potenziare i processi decisionali tipicamente connessi alle scelte di costituzione dell'impresa.
Viene qui proposto un metodo d'indagine che può dimostrarsi utile per individuare benchmark strategici, altamente flessibile e adattabile a casi specifici.

In particolare, il presente lavoro si articola come segue.

Nel capitolo 1 si presenta una panoramica delle analisi di settore nelle diverse discipline economiche.
Nel capitolo 2 si fornisce una descrizione generale degli algoritmi di clustering e delle loro comuni applicazioni.
Nel capitolo 3 si presentano nel dettaglio la metodologia proposta e gli strumenti impiegati per individuare e analizzare i cluster in cui tende a segmentarsi il settore oggetto di studio.
Nel capitolo 4 si presenta un caso d'uso su una popolazione di imprese, appartenenti a un medesimo settore, e qualificate da una serie di variabili di natura contabile che ne descrivono la strategia di investimento.
Con tale esempio, si vuole identificare la strategia di investimento tipicamente più adottata dalle imprese "di benchmark", per comprendere, data una serie di voci di bilancio selezionate, quanto conviene investire in ciascuna di esse.
Nel capitolo 5, infine, si presenta un secondo esempio, focalizzato sull'analisi dei cluster delle province italiane, qualificate da diverse informazioni sulla filiera in cui un dato settore obiettivo si inserisce.
Lo scopo di questa seconda analisi consiste nell'individuazione delle aree geografiche in cui può esservi un insieme di condizioni ambientali particolarmente favorevoli all'ingresso di una nuova azienda operante in tale settore.
Nelle conclusioni si traggono i principali elementi innovativi e si discutono i principali limiti di una ricerca di settore fondata sul metodo presentato, e si propongono ulteriori sviluppi che consentirebbero di sfruttare ulteriormente le potenzialità economico-gestionali del clustering.

Pur cambiando dataset, aventi come osservazioni imprese, aree geografiche o interi settori, la metodologia di base rimane invariata, e può essere estesa a numerosi problemi di benchmarking e di definizione strategica in ambito economico.

2. Il concetto di settore

In questa sede il significato attribuito al termine "settore", o "industria", coincide con quello più generico: un insieme di imprese che svolgono attività economiche volte alla produzione di beni e servizi con un certo grado di omogeneità.
Non deve essere quindi inteso secondo la rigida distinzione tra i comparti primario, secondario e terziario, né secondo quella tra settore pubblico e privato o tra settore manifatturiero e di servizi. Non si fa qui neanche riferimento ai c.d. studi di settore, considerati in materia fiscale per calcolare i parametri di reddito dei contribuenti[7].
La classificazione industriale può seguire un certo grado di arbitrarietà. Ad esempio, Confindustria identifica 24 macrocategorie di settori, e per ciascuna di queste una serie di sottocategorie[8].
Diversi istituti finanziari, statistici, governativi e di ricerca propongono una propria classificazione merceologica[9]. La classificazione ufficiale e più impiegata in Italia è la classificazione Ateco (ATtività ECOnomiche) proposta dall' Istat per la prima volta nel 2002 e poi aggiornata nel 2007. In Europa esiste la classificazione NACE (Nomenclature statistique des Activités économiques dans la Communauté Européenne) istituita dall'Eurostat nel 1970. Negli Stati Uniti, in Canada e in Messico vengono invece adottati i codici NAICS (North American Industry Classification System), che nel 1997 hanno sostituito il precedente sistema SIC (Standard Industrial Classification) del 1937.
Le prime cifre di ciascun codice rappresentano delle macrocategorie settoriali, mentre le successive identificano una subcategoria sempre più definita.
Naturalmente al crescere della precisione definitoria di un'attività economica si riduce il numero di imprese che la esercita, e questo fattore pone un dilemma importante in materia di analisi dati, dove la significatività dei risultati dipende da una sufficiente numerosità delle osservazioni.
In questa analisi si adotterà la classificazione Ateco come punto di riferimento nella definizione dei dataset di settore, di filiera e di provincia.
Diversi autori effettuano analisi di settore senza utilizzare particolari strumenti di analisi dati, ma affidandosi in primo luogo a censimenti governativi (come i censimenti dell'Istat) e ai valori medi delle imprese identificate da un determinato codice merceologico.
Ad esempio, Vitali[10] svolge un'analisi qualitativa dei distretti industriali piemontesi facendo ricorso ai codici Ateco per distinguere le attività produttive.
È tuttavia possibile identificare le aziende appartenenti a un dato settore in base ad altri criteri, come le classificazioni merceologiche o liste autonomamente redatte di imprese produttrici di particolari beni o servizi. È anche possibile estendere l'analisi a imprese operanti al di fuori dell'Italia o effettuare confronti a livello di paese, facendo ricorso ad altre classificazioni industriali.

---

[7] Enciclopedia Treccani - Dizionario di Economia e Finanza (2012) voce "settore"
Boggia, A., Carucci, A. M. M., & Filippello, R. istat working papers.
[8] Enciclopedia Treccani - Dizionario di Economia e Finanza (2012) voce "settore - settore industriale"
[9] Dalziel, M. (2007). A systems-based approach to industry classification. *Research Policy*, *36*(10), 1559-1574.
Schnabl, E., & Zenker, A. (2013). *Statistical classification of knowledge-intensive business services (KIBS) with NACE Rev. 2* (Vol. 25). Karlsruhe: Fraunhofer ISI.
[10] di ricerca Ceris-Cnr, G., Calabrese, G., Corio, G., Finardi, U., Manello, A., Ragazzi, E., ... & Saracco, P. Le caratteristiche socio-economiche dei cluster di imprese in Piemonte.

2.1 L'influenza della filiera nella letteratura economica

Fra le condizioni ambientali che maggiormente incidono sulle caratteristiche di un settore, centrale è il ruolo della filiera produttiva, o supply chain, che connette secondo schemi comuni le imprese appartenenti a settori diversi.

Ai fini dell'analisi strategica la filiera può essere definita come l'insieme degli attori (delle loro caratteristiche e delle loro strategie) che intervengono nel ciclo di trasformazione fisico-tecnica ed economica di un bene, giocando un ruolo nel processo di formazione del valore[11].

I rapporti fra più imprese appartenenti a settori diversi possono consistere in rapporti verticali o orizzontali. Nel caso di rapporti verticali, si tratta di rapporti di fornitura, rispetto ai quali l'impresa può posizionarsi a monte o a valle[12]. Nel caso dei rapporti orizzontali, questi possono consistere in rapporti di concorrenza o di collaborazione, a seconda che i beni o i servizi prodotti nei due settori siano da qualificarsi come beni sostituti o complementari[13].

I rapporti tra le imprese appartenenti alla medesima filiera produttiva si strutturano in senso verticale, mentre si collegano in senso orizzontale tutte le imprese che operano in uno stesso stadio del ciclo produttivo.

Nel management, la posizione di un settore rispetto alla sua supply chain viene studiata per mezzo del Modello delle cinque forze competitive di Porter[14].

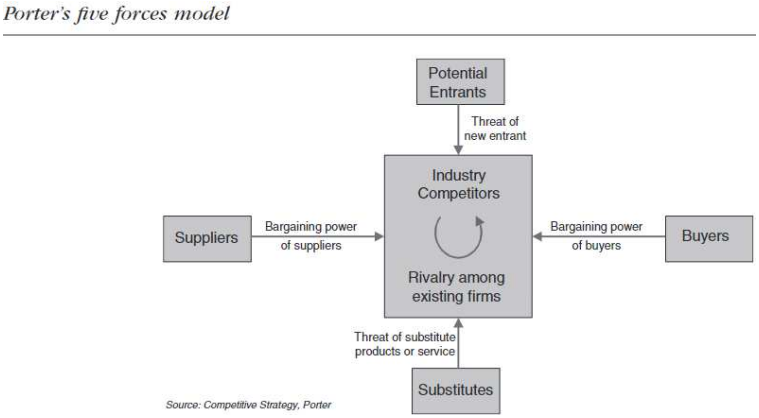


FIGURA 1 MODELLO DELLE CINQUE FORZE DI PORTER

Il concetto di filiera produttiva e l'impatto che questa produce sul singolo settore è centrale nello studio di Porter, dove assumono risalto il potere contrattuale dei fornitori e il potere contrattuale dei compratori[15]. Il modello di Porter attribuisce anche molta importanza alle barriere all'entrata presenti nel settore, che garantiscono una stabilità maggiore alle imprese già operanti al suo interno, limitando l'ingresso di nuove imprese nel mercato.

[11] Enciclopedia Treccani - Dizionario di Economia e Finanza (2012) voce "filiera"
[12] Zamora, E. A. (2016). Value chain analysis: A brief review. *Asian Journal of Innovation and Policy*, *5*(2), 116-128.
[13] Musso, F. (2012). *Innovazione nei canali di marketing*. Clueb.
[14] Grundy, T. (2006). Rethinking and reinventing Michael Porter's five forces model. *Strategic change*, *15*(5), 213-229.
[15] Skjott-Larsen, T. (2007). *Managing the global supply chain*. Copenhagen Business School Press DK.

In Italia, diverse filiere produttive importanti assumono la forma di distretti industriali, indagati dall'Istat attraverso la costituzione dei SLL (Sistemi Locali del Lavoro)[16].

Un distretto industriale è un sistema produttivo costituito da un insieme di imprese, prevalentemente di piccole e medie dimensioni, caratterizzate da una tendenza all'integrazione orizzontale e verticale e alla specializzazione produttiva, in genere concentrate in un determinato territorio e legate da una comune esperienza storica, sociale, economica e culturale[17].

[16] Sforzi, F. (Ed.). (1997). *I sistemi locali del lavoro 1991*. Roma: Istat.
Coppola, G., & Mazzotta, F. (2005). I sistemi locali del Lavoro in Italia: Aspetti teorici ed empirici.
Calafati, A. G., & Compagnucci, F. (2005). Oltre i sistemi locali del lavoro. *Economia Marche*, *24*(1), 51-76.

[17] Enciclopedia Treccani - Dizionario di Economia e Finanza (2012) voce "distretto industriale"

2.2 Le analisi di settore nella letteratura economica

Le analisi di settore sono una fase chiave del management e del decision making di qualsiasi impresa di qualsiasi età e dimensione.
Le analisi di settore consentono, infatti, di conoscere la concorrenza, sviluppare nuovi prodotti, segmentare temporalmente e geograficamente il mercato, decidere in merito a fusioni e acquisizioni aziendali, redigere piani di budget e investimento, e, più in generale, individuare i pattern decisionali e le condizioni ambientali che hanno fatto emergere le imprese di successo e che hanno generato inefficienze nelle imprese in declino.
La dimensione ormai globale e iperconnessa dei mercati impone, a chiunque sia chiamato a prendere decisioni in merito alla strategia aziendale, di tenere in considerazione l'ambiente esterno dell'impresa. Possono essere d'esempio il caso della recente crisi dei microchip che ha colpito il settore automotive[18], il monopolio che ha mantenuto la Fiat per i suoi fornitori[19] e i recenti sviluppi dell'IOT (Internet Of Things) e delle blockchain per i quali assumono sempre più rilevanza i rapporti tra aziende del settore manifatturiero e produttori di software[20].
Le analisi di settore sono altresì un tema multidisciplinare dell'economia.
Diversi autori, quando effettuano un'analisi di settore, si concentrano maggiormente sugli aspetti qualitativi del settore, come le barriere legali, la strategia d'impresa, la cultura aziendale e le risorse sul territorio. La nostra intenzione è quella di identificare una metodologia quantitativa accessibile per il singolo. Di seguito vengono citate solo alcune delle metodologie proposte in materia economica e al di fuori del campo statistico. In particolare, si riportano gli strumenti a tal fine più indicati in materia di bilancio, nel management e in macroeconomia.

In ambito contabile, Navaroni[21] presenta la Position Analysis come la più evoluta tra le analisi di bilancio:
"si definiscono analisi di posizione, o position analysis, quelle volte a confrontare gli indici, calcolati sul bilancio di un'impresa, con gli indici medi di settore nel quale l'impresa si trova a operare."
Nella Position Analysis, gli indici di performance di un'impresa coincidono con gli indici di analisi di bilancio, come il ROI, il ROS o il Turnover Ratio. Si tratta della sola tipologia di analisi di bilancio che tiene in considerazione lo stato di salute del settore in cui l'impresa si inserisce.

[18] Russo, M., Alboni, F., Sanginés, J. C., De Domenico, M., Mangioni, G., Righi, S., & Simonazzi, A. (2022). The Changing Shape of the World Automobile Industry: A Multilayer Network Analysis of International Trade in Components and Parts. *Institute for New Economic Thinking Working Paper Series*, (173).
[19] Balcet, G., & Enrietti, A. (1997). Regionalisation and globalisation in Europe: The case of Fiat Auto Poland and its suppliers. *Les Actes de GERPISA*, (20).
[20] Malik, S., Dedeoglu, V., Kanhere, S. S., & Jurdak, R. (2019, July). Trustchain: Trust management in blockchain and iot supported supply chains. In *2019 IEEE International Conference on Blockchain (Blockchain)* (pp. 184-193). IEEE.
Pal, K. (2023). Internet of Things Impact on Supply Chain Management. *Procedia Computer Science*, *220*, 478-485.
[21] Navaroni, M. (2020). La Position Analysis: la più Evoluta tra le Analisi di Bilancio. *Economia Aziendale Online-*, *11*(2), 133-144.

Nel management, Aithal[22], ideatore dell'analisi ABCD[23], presenta una panoramica dei metodi di analisi di settore più diffusi in materia.
Una descrizione sintetica della lista da lui proposta è presentata di seguito, ma si rimanda alla bibliografia per una definizione più dettagliata.

Types of Industry Analysis:

- Industry Sector Analysis: analisi delle diverse sottospecie di attività presenti in ciascun settore
- Industry Trend Analysis: analisi previsiva delle performance future di tutte le aziende del settore
- Environmental Analysis: analisi dei fattori ambientali che condizionano l'industria
- Competitor Analysis: analisi delle strategie adottate nel settore per fronteggiare la concorrenza
- Alternative Product/service Analysis: analisi dei modelli di business che caratterizzano l'offerta del settore
- Financial Performance Analysis: analisi dei valori medi di bilancio delle imprese di un dato settore (analoga alla Position Analysis precedentemente descritta)
- Industry ABCD Analysis: analisi dei vantaggi, dei benefici, dei vincoli e degli svantaggi di un settore
- Industry SWOT Analysis: analisi dei punti di forza, di debolezza, delle opportunità e delle minacce del settore
- Product/Service Analysis: analisi della qualità dei prodotti e servizi del settore
- Investment Analysis: analisi dei livelli di investimento nello stock market del settore
- Automation & Labour Requirement Analysis: analisi dell'impatto che i sistemi di automazione adottabili in un settore hanno sul mercato del lavoro del settore medesimo.
- People perception Analysis: analisi della percezione che il pubblico ha nei confronti dell'utilità e dell'impatto sociale del settore.
- Size of the Industry & Total Contribution to the Economy: analisi dell'impatto del settore sul sistema economico nel suo complesso e sulla quota di PIL prodotta.
- Market Demand Analysis: analisi dei volumi della domanda dei consumatori
- Opportunity Analysis: analisi delle opportunità accessibili nel settore
- Government Policy Analysis: analisi delle politiche nazionali sulla profittabilità del settore
- Industry Contribution & Employment Generation Analysis: analisi dell'impatto positivo che il settore ha sugli altri settori e sul tasso di occupazione complessivo
- Top leading Companies in an Industry & their Strategies: analisi delle imprese leader del settore
- Latest Industrial Developments: analisi degli impatti che le moderne tecnologie possono produrre sul settore

[22] Aithal, P. S. (2017). Industry Analysis–The First Step in Business Management Scholarly Research. *International Journal of Case Studies in Business, IT and Education (IJCSBE)*, *1*(1), 1-13.
[23] Aithal, P. S. (2017). ABCD Analysis as Research Methodology in Company Case Studies. *International Journal of Management, Technology, and Social Sciences (IJMTS)*, *2*(2), 40-54.

- Cross-industry Analysis: analisi della filiera produttiva e dei comuni legami verticali e orizzontali delle imprese del settore con le imprese degli altri settori
- Market Size: analisi degli andamenti di crescita dimensionale del settore
- Information Technology Implementation: analisi del livello di ICT nelle diverse fasi di sviluppo del settore
- Studying Industry Innovations using Six Thinking Hats: analisi del settore per mezzo della metodologia proposta da Edward de Bono nel 1985

In macroeconomia e politica economica, le analisi di settore vengono effettuate attraverso complessi modelli econometrici, definiti matrici Input Output o matrici di contabilità nazionale (NAM)[24]. La contabilità nazionale (o contabilità macroeconomica) si propone di descrivere quantitativamente, in termini monetari, l'attività economica e finanziaria di un paese mediante un insieme coerente di conti e identità[25].
Una tavola input output, secondo lo schema più basilare, riporta per ogni coppia di settori il valore monetario dell'output del primo settore consumato come input del secondo settore[26].
Tramite questi strumenti, è possibile controllare lo stato di salute del singolo settore, delle filiere produttive, delle supply chain, del sistema economico nel suo complesso o a livello sub-nazionale.
I principali ideatori delle matrici di contabilità nazionale sono stati Leontief, per le IOT (input-output table) nel 1936 e Stone per le SAM (social accounting matrix) nel 1961.
Secondo Round[27], la SAM è una matrice quadrata, comprensiva e flessibile. È quadrata, perché riporta le transazioni tra tutti i settori economici, che sono riportati sia sulle righe sia sulle colonne, e ogni cella riporta il valore economico consumato dal settore di input (sulla riga) e prodotto dal settore di output (sulla colonna). È comprensiva, perché riporta tutte le dimensioni di consumo, produzione, accumulo e distribuzione presenti nel sistema economico. È flessibile, perché può essere costruita su diversi livelli di disaggregazione, per rappresentare e studiare il sistema nel suo complesso a diversi gradi di dettaglio.
Nello stesso ambito l'Istat, la Banca d'Italia e la maggior parte delle istituzioni finanziarie e governative europee fa riferimento al modello Sec[28] (Sistema europeo dei conti nazionali e regionali), creato a partire dal modello Sna[29] (una SAM), adottato negli Stati Uniti. Dei modelli Sec e Sna sono state sviluppate più versioni, come lo Sna53, lo Sna68, il Sec70, il Sec95, lo Sna2008 e molti altri. Mentre il sistema Sna classifica le attività produttive in base ai codici NAICS, il sistema Sec utilizza i codici NACE. Una matrice di contabilità nazionale per essere costruita richiede una mole di dati inaccessibile per il singolo, ma diversi istituti finanziari e governativi rendono disponibili alcune delle matrici che redigono. Un esempio può essere quella pubblicata sul sito dell'Istat[30].

[24] Saieva, V. (2012). International production relocation. *Economic Focus,(2)*, 1-26.
Nicolardi, V., & Marini, C. (2016). L'aggiornamento di strutture di Contabilità Nazionale disaggregate. In *Metodi e Analisi statistiche 2016* (pp. 131-146). Università degli Studi di Bari" Aldo Moro".
[25] Treccani - Enciclopedia delle scienze sociali (1992) voce "contabilità nazionale" - introduzione
[26] TANAKA, F. J. (2011). Applications of Leontief's input-output analysis in our economy.
[27] Round, J. (2003). Social accounting matrices and SAM-based multiplier analysis. *The impact of economic policies on poverty and income distribution: Evaluation techniques and tools*, *14*, 261-276.
[28] Giovanelli, L. (Ed.). (2006). *Contabilità dello stato e sistema europeo dei conti (SEC95) nella prospettiva comunitaria*. Giuffrè Editore.
[29] Santos, S. (2011). Constructing SAMs from the SNA.
[30] Mastrantonio (2018) https://www.istat.it/it/archivio/209141

Oltre a quelle qui riportate, esistono molte altre metodologie, di natura quantitativa e qualitativa, per svolgere l'analisi di un settore specifico. Non è quindi presente una procedura univocamente o prevalentemente adottata dagli economisti per svolgere analisi di settore, e non vi è nemmeno una definizione univoca di cosa sia un'analisi di settore. Si può quindi considerare l'idea di adottare metodi statistici, o econometrici, per raggiungere gli obiettivi conoscitivi generalmente proposti in tale analisi.

2.3 L'adozione dei cluster nelle analisi di settore

Molto spesso, nella letteratura economica, si parla di cluster di imprese. Non sempre il significato assunto da questo termine coincide con quello che vi è normalmente attribuito in campo statistico (si vedano ad esempio, fra i tanti, Parente, Smith e Kim[31]).
Il termine "cluster" significa semplicemente "gruppo" e viene adottato da analisti giuridici, finanziari e contabili per indicare una qualsiasi selezione di imprese accomunate da una certa caratteristica. Spesso si parla di cluster di imprese con riferimento a quelle appartenenti a una particolare supply chain, a un distretto industriale o a un SLL (sistema locale del lavoro), oppure a quelle accomunate dal fatto di essere grandi imprese o PMI (piccole e medie imprese) operanti in una data regione geografica, impegnate nel medesimo settore o guidate dalla medesima holding.
In statistica il termine ha una connotazione molto più specifica e quantitativa.
Un cluster è un sottoinsieme di osservazioni presenti in una popolazione relativamente simili fra loro. Secondo Skiena[32], il clustering è il problema di raggruppare punti per somiglianza.
Tale concetto di somiglianza viene assimilato al concetto di distanza, in modo tale da poter essere misurato quantitativamente. Ogni caratteristica considerata nella popolazione, e quindi misurata attraverso una variabile, può essere trattata come una coordinata nello spazio. Il problema di misurare la dissimilarità fra due osservazioni, in questo caso tra due imprese, si traduce quindi nel problema di calcolare la distanza tra due punti.
Nel prossimo capitolo viene proposta una panoramica dei principali algoritmi di clustering e delle diverse metriche di distanza generalmente adottate.

[31] Parente, R. (2008). Co-evoluzione e cluster tecnologici. Roma: Aracne.
Smith, R. V. (2003). Industry cluster analysis: Inspiring a common strategy for community development. *Central Pennsylvania Workforce Development Corporation Report*, *296*.
Kim, H., Hwang, S. J., & Yoon, W. (2023). Industry cluster, organizational diversity, and innovation. *International Journal of Innovation Studies*, *7*(3), 187-195.
[32] Skiena, S. S. (2017). *The data science design manual* Cap. 10.5: clustering. Springer.

Attraverso questa metodologia è possibile studiare dataset contenenti numerose informazioni di bilancio di un gran numero di imprese e ricondurli a poche statistiche descrittive e visualizzazioni sintetiche. Si può così conoscere il modo in cui una popolazione di imprese, come quelle appartenenti a un settore, si stratifica, e le caratteristiche di ciascuna delle fasce che la compongono.
L'idea di adottare il clustering per effettuare analisi di settore è stata attuata da diversi autori, in Italia e all'estero.
Curea et al[33] utilizzano il clustering per studiare le performance del settore minerario dei paesi dell'Europa dell'Est negli anni della crisi del 2008, avvalendosi dei codici NACE e il database Orbis di BvD (Bureau van Djick).
Abbas et al[34] utilizzano il clustering per analizzare il settore manifatturiero indonesiano.
Russo et al[35] utilizzano il clustering per analizzare i distretti industriali del settore meccanico italiano, avvalendosi dei SLL.
Fanelli et al[36] utilizzano il clustering, i codici ATECO e il database Aida di BvD (Bureau van Djick) per studiare i cluster dei birrifici italiani.

[33] Curea, S. C., Belascu, L., & Barsan, A. M. (2020). An Exploratory Study of Financial Performance in CEE Countries. *KnE Social Sciences*, 286-300.
[34] Abbas, A., Prayitno, P., Nurkim, N., Prumanto, D., Dewadi, F. M., Hidayati, N., & Windarto, A. P. (2021, February). Implementation of clustering unsupervised learning using K-Means mapping techniques. In *IOP Conference Series: Materials Science and Engineering* (Vol. 1088, No. 1, p. 012004). IOP Publishing.
[35] Russo, M., Pirani, E., & Paterlini, S. (2006). L'industria meccanica in Italia: una analisi cluster delle differenze territoriali. *Materiali di discussione*.
[36] Fanelli, R. M., & Felice, F. (2014). Un'applicazione dell'analisi multivariata e della convergenza non parametrica all'industria birraria italiana. *Italian Review of Agricultural Economics*, *69*(1), 7-30.

3. Tassonomia degli algoritmi di clustering

Gli algoritmi di clustering consentono di ridurre la complessità di dataset comprendenti una vasta ed eterogenea popolazione campionaria, riconducendo tutte le osservazioni alle statistiche descrittive di pochi gruppi omogenei[37].

I modelli di clustering rientrano nella macrocategoria dei modelli di unsupervised learning, in quanto producono risposte in assenza di dati di addestramento etichettati, ovvero in assenza di variabili di output di esempio. Vengono così distinti dai modelli di supervised learning, come le analisi discriminanti e le regressioni, nei cui dati di addestramento è presente un esempio di risposta, o output, per ciascuna osservazione di input.

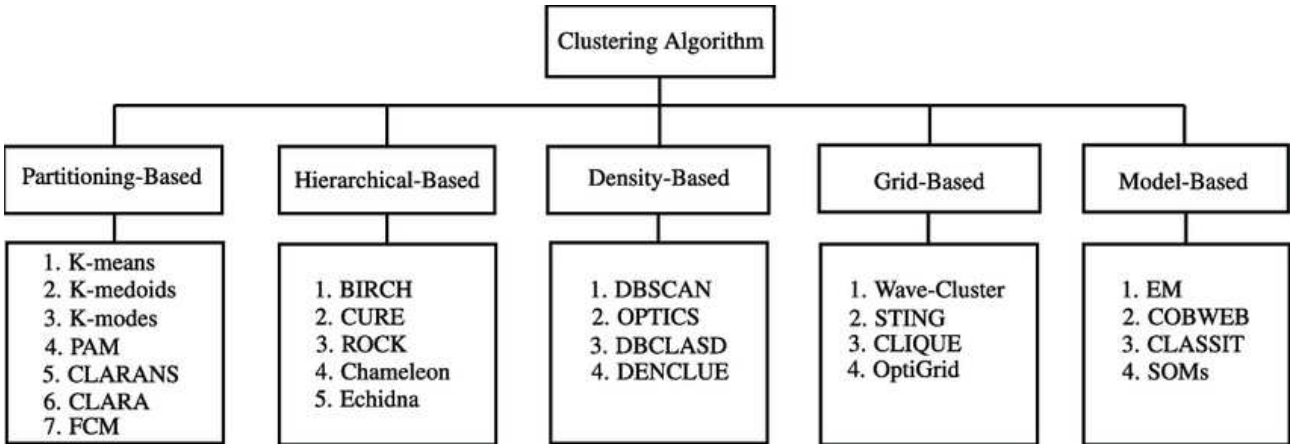


FIGURA 2 CLASSIFICAZIONE DEGLI ALGORITMI DI CLUSTERING[38]

Tra gli algoritmi di clustering sono poi presenti molte altre distinzioni, come quella gli algoritmi di clustering partizionale e quelli di clustering gerarchico, le due categorie più semplici e generalmente più adottate[39], alle quali si dedicano alcuni dei paragrafi successivi.

In sintesi, il clustering partizionale[40] predilige la convergenza delle osservazioni. Si inizializzano tanti centri quanti sono i cluster nella popolazione, e li si correggono ad ogni iterazione in modo tale da minimizzare la dispersione dei gruppi. L'algoritmo di clustering partizionale, e in generale di clustering, più diffuso è probabilmente il kmeans. Si tratta di un algoritmo estremamente semplice, con buone performance e un bassissimo costo computazionale.

Il clustering gerarchico, invece, si costruisce direttamente a partire dalle distanze tra ogni coppia di osservazioni, e unisce (o separa), progressivamente, le più vicine (o le più lontane).

[37] Fung, G. (2001). A comprehensive overview of basic clustering algorithms.

[38] Fahad, A., Alshatri, N., Tari, Z., Alamri, A., Khalil, I., Zomaya, A. Y., ... & Bouras, A. (2014). A survey of clustering algorithms for big data: Taxonomy and empirical analysis. *IEEE transactions on emerging topics in computing*, *2*(3), 267-279.

[39] Sonagara, D., & Badheka, S. (2014). Comparison of basic clustering algorithms. *Int. J. Comput. Sci. Mob. Comput*, *3*(10), 58-61.

[40] Li, Y., & Wu, H. (2012). A clustering method based on K-means algorithm. *Physics Procedia*, *25*, 1104-1109.

Tra le altre categorie di clustering si ricordano anche il clustering di densità e di griglia.
Il clustering basato sul concetto di densità aggrega insieme i gruppi di punti più densi, e non richiede di specificare il numero di gruppi[41]. Ne è esempio il DBSCAN[42], che utilizzeremo per rimuovere gli outlier, e che consente di identificare gruppi con forme anomale.
Il clustering di griglia[43] si fonda sul concetto di divisione dello spazio in aree, o celle, sulla base delle posizioni occupate dalla maggior parte dei punti. Si aggregano poi i punti posizionati nelle aree assegnate a un medesimo cluster.
Esistono poi numerosi sistemi di clustering più avanzati e complessi, basati su dei modelli di apprendimento automatico. Un esempio possono essere le reti neurali SOM[44] (Self Organizing Maps o Mappe di Kohonen) le cui celle, o neuroni, modificano i propri valori in modo tale da replicare il più possibile i valori medi di una particolare fascia del campione.

Data la grande varietà di metodi, più o meno diffusi, risulta complesso decidere quale adottare nel caso specifico.
La scelta dell'algoritmo di clustering può dipendere da[45]:

1. semplicità di calcolo e costo computazionale
2. indici di valutazione dei cluster, come la silhouette o gli indici di Dunn, DB o CH
3. conoscenza del dominio del fenomeno studiato
4. le dimensioni dei dati: certi algoritmi funzionano bene per piccoli dataset (come il clustering gerarchico) mentre altri per grandi dataset (come le SOM)
5. tipologia di task tipicamente risolto tramite una determinata tipologia di algoritmo

In questa analisi utilizzeremo il kmeans, il kmedoids (o PAM, Partition Around Medoids), e il DBSCAN. Utilizzeremo il DBSCAN per rimuovere gli outlier, il kmeans per individuare il numero ottimale di cluster e il kmedoids per produrre cluster stabili e di dimensioni omogenee.
Nei prossimi paragrafi verranno descritti alcuni algoritmi di clustering comunemente utilizzati, nonché degli strumenti solitamente complementari al loro utilizzo.

[41] Batra, P. (2018). Comparative study of density based clustering algorithms.
Jahirabadkar, S., & Kulkarni, P. (2014). Algorithm to determine ε-distance parameter in density based clustering. *Expert systems with applications*, *41*(6), 2939-2946.
[42] Hahsler, M., Piekenbrock, M., & Doran, D. (2019). dbscan: Fast density-based clustering with R. *Journal of Statistical Software*, *91*, 1-30.
[43] Cheng, W., Wang, W., & Batista, S. (2018). Grid-based clustering. In *Data clustering* (pp. 128-148). Chapman and Hall/CRC.
Hireche, C., Drias, H., & Moulai, H. (2020). Grid based clustering for satisfiability solving. *Applied Soft Computing*, *88*, 106069.
[44] Gelbard, R., Goldman, O., & Spiegler, I. (2007). Investigating diversity of clustering methods: An empirical comparison. *Data & Knowledge Engineering*, *63*(1), 155-166.
[45] Rodriguez, M. Z., Comin, C. H., Casanova, D., Bruno, O. M., Amancio, D. R., Costa, L. D. F., & Rodrigues, F. A. (2019). Clustering algorithms: A comparative approach. *PloS one*, *14*(1), e0210236.
https://generativeai.pub/how-to-choose-the-right-clustering-algorithm-for-your-data-8f3ee24b9c16

3.1 La matrice delle distanze

Ogni algoritmo di clustering si basa sul concetto di distanza: ogni variabile rappresenta una coordinata nello spazio, e ogni osservazione è un punto che assume una certa posizione[46]. Pertanto, è possibile misurare la distanza che separa due qualsiasi osservazioni nei dati.
Esistono molte metriche per il calcolo delle distanze. Le più ricorrenti sono la Distanza Euclidea, la Distanza di Manhattan (anche detta distanza dei taxi) e le Distanze di Chebyshev (Maximum per $p = +\infty$ e Minimum per $p = -\infty$), che possono essere tutte rappresentate come casi particolari della più generale Distanza di Minkowski[47]. Quest'ultima può essere definita nel modo seguente per ogni coppia di osservazioni *i, j* rispetto a una serie di variabili $l = 1 \ldots d$:

$$D(\mathbf{x}_i, \mathbf{x}_j) = \left(\sum_{l=1}^{d} |x_{il} - x_{jl}|^{1/p} \right)^p$$

p	Distance measure
p=1	Manhattan
p=2	Euclidean
p→∞	Chebyshev

In particolare, la distanza di Chebyshev può assumere le forme:
$D(x_i, x_j) = \max(|x_{il} - x_{jl}|)\ per\ \ p = +\infty$
$D(x_i, x_j) = \min(|x_{il} - x_{jl}|)\ \ per\ \ p = -\infty$

Va precisato che, in presenza di variabili normalizzate e che condividono la medesima distribuzione, la distanza di Manhattan si presenta come la più insensibile agli outlier.
Nel clustering, ogni operazione può essere effettuata a partire da una matrice delle distanze.
Una matrice delle distanze è una matrice quadrata simmetrica, contenente, per ogni coppia di osservazioni, la relativa distanza. La matrice ha solo valori nulli sulla diagonale, in quanto la distanza tra ogni osservazione e la medesima è sempre pari a 0 (ovvero per $i = j, D(x_i, x_j) = 0$).

```
> dataset
     var1 var2 var3
obs1   -1   -2    0
obs2    1   -2   -1
obs3    0    0    0
obs4    0   -1    0
> as.matrix(dist(dataset, method='manhattan'))
     obs1 obs2 obs3 obs4
obs1    0    3    3    2
obs2    3    0    4    3
obs3    3    4    0    1
obs4    2    3    1    0
```

FIGURA 3 ESEMPIO DI DATASET E RELATIVA MATRICE DELLE DISTANZE CON LA METRICA DI MANHATTAN

La matrice consente anche di valutare la bontà del clustering effettuato, e di calcolare indici di valutazione dei cluster come la silhouette.
Graficamente, è possibile rappresentare una matrice delle distanze con una heatmap (le cui ricche possono essere ordinate attraverso un dendrogramma[48]) o con un grafo (si vedano le figure 24 e 41).

[46] Skiena, S. S. (2017). *The data science design manual* Cap. 10.1: measuring distances. Springer.
[47] Rodrigues, É. O. (2018). Combining Minkowski and Chebyshev: New distance proposal and survey of distance metrics using k-nearest neighbours classifier. *Pattern Recognition Letters*, *110*, 66-71.
Singh, A., Yadav, A., & Rana, A. (2013). K-means with Three different Distance Metrics. *International Journal of Computer Applications*, *67*(10).
[48] Tsafrir, D., Tsafrir, I., Ein-Dor, L., Zuk, O., Notterman, D. A., & Domany, E. (2005). Sorting points into neighborhoods (SPIN): data analysis and visualization by ordering distance matrices. *Bioinformatics*, *21*(10), 2301-2308.

3.2 Il clustering gerarchico

Il clustering gerarchico è stato introdotto da Sokal e Sneath[49] nel 1958, come uno strumento per l'analisi dei dati biologici, e rappresenta il primo algoritmo di clustering della storia.
Nel clustering gerarchico si distinguono gli algoritmi top down da quelli bottom up[50].
Mentre gli algoritmi top down (o divisivi) partono dall'alto, raccogliendo inizialmente tutte le osservazioni in un unico gruppo, e cercando ad ogni iterazione di dividere i dati per produrne uno nuovo, gli algoritmi bottom up (o agglomerativi) partono dal basso, mantenendo inizialmente separate tutte le osservazioni, per poi aggregare la coppia più simile ad ogni step.
La somiglianza tra più gruppi di osservazioni può essere calcolata secondo diversi metodi di "linkage"[51], di solito raccogliendo la distanza singola, completa o media tra un gruppo e un altro.
La distanza media è la media delle distanze tutte le coppie di elementi nei due gruppi, la distanza completa è la massima e la distanza singola è la minima.
I metodi di linkage non vanno confusi con le metriche di distanza illustrate nel precedente capitolo, che misurano la similarità tra due osservazioni singole.

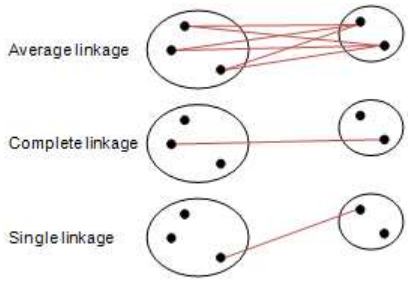


FIGURA 4 DIFFERENZE TRA I METODI DI LINKAGE PER IL CALCOLO DELLA DISTANZA TRA GRUPPI[52]

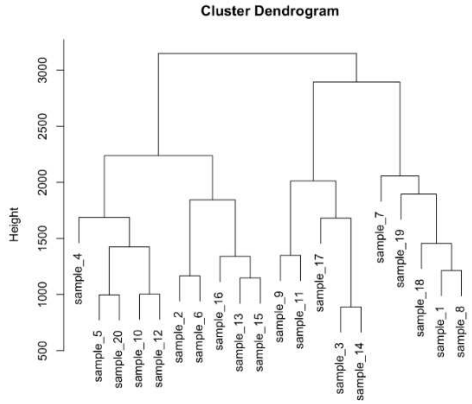


FIGURA 5 ESEMPIO DI DENDROGRAMMA[53]

[49] Sokal, R. R., & Sneath, P. H. (1963). Principles of numerical taxonomy. *Principles of numerical taxonomy*. Ehrlich, P. R. (1958). Problems of higher classification. *Systematic Zoology*, *7*(4), 180-184.
[50] Shetty, P., & Singh, S. (2021). Hierarchical clustering: a survey. *International Journal of Applied Research*, *7*(4), 178-181.
[51] Yim, O., & Ramdeen, K. T. (2015). Hierarchical cluster analysis: comparison of three linkage measures and application to psychological data. *The quantitative methods for psychology*, *11*(1), 8-21.
[52] Mahmoud, M. (2012). *Genotype imputation based on discriminant and cluster analysis* (Master's thesis, Norwegian University of Life Sciences, Ås).
[53] https://r-graph-gallery.com/29-basic-dendrogram.html

Il clustering gerarchico può essere graficamente rappresentato mediante un dendrogramma[54] (un albero costituito da dei nodi, o rami, rappresentanti un particolare sottogruppo a un determinato step, e da delle foglie, rappresentanti le singole osservazioni) nel quale si può quindi identificare una gerarchia dei passaggi divisivi o aggregativi.
Per costituire dei cluster attraverso il dendrogramma, è necessario "tagliare l'albero"[55], in base al numero di gruppi che si vogliono costituire (ovvero di rami che si lasciano al di sopra del taglio) o all' "altezza dell'albero", ovvero al grado di dispersione complessiva che si può accettare.
Il dendrogramma consente anche di ordinare le righe di una matrice delle distanze, in modo tale da concentrare su posizioni vicine osservazioni simili.

3.3 Il clustering partizionale

Il clustering partizionale si fonda sul concetto di centri: ogni osservazione è assegnata al gruppo con il centro più vicino, e ad ogni iterazione viene corretto e riposizionato il centro di ciascun gruppo, in modo tale da rappresentarlo meglio.
Gli algoritmi più ricorrenti appartenenti a questa categoria sono due: il kmeans e il kmedoids[56].
Il kmeans è stato proposto da MacQueen[57] nel 1967, come una generalizzazione del metodo delle medie aritmetiche per l'analisi dei dati multivariati. Il kmedoids è stato sviluppato da Kaufman e Rousseeuw[58] nel 1987, come un metodo più robusto e meno sensibile agli outlier rispetto al kmeans. Gli stessi Kaufman e Rousseeuw hanno proposto l'algoritmo Clara, che utilizza campioni casuali per ridurre il costo computazionale del kmedoids.

L'algoritmo kmeans trova i centroidi dei cluster, mentre l'algoritmo kmedoids i medoidi.
L'algoritmo kmeans crea dei punti artificiali ad ogni iterazione, le cui coordinate sono calcolate come la media delle coordinate dei punti di ciascun cluster, mentre il kmedoids seleziona dei punti esistenti nel dataset, individuando ad ogni step dei nuovi centri che minimizzano la dispersione complessiva dei cluster maggiormente rispetto ai precedenti.
I due algoritmi procedono dunque nelle seguenti fasi[59]:

---

[54] Forina, M., Armanino, C., & Raggio, V. (2002). Clustering with dendrograms on interpretation variables. *Analytica Chimica Acta*, *454*(1), 13-19.
[55] Boudaillier, E., & Hebrail, G. (1998). Interactive interpretation of hierarchical clustering. *Intelligent Data Analysis*, *2*(1-4), 229-244.
[56] Arbin, N., Suhaimi, N. S., Mokhtar, N. Z., & Othman, Z. (2015, December). Comparative analysis between k-means and k-medoids for statistical clustering. In *2015 3rd International Conference on Artificial Intelligence, Modelling and Simulation (AIMS)* (pp. 117-121). IEEE.
[57] MacQueen, J. (1967, June). Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability* (Vol. 1, No. 14, pp. 281-297).
[58] Schubert, E., & Rousseeuw, P. J. (2021). Fast and eager k-medoids clustering: O (k) runtime improvement of the PAM, CLARA, and CLARANS algorithms. *Information Systems*, *101*, 101804.
[59] Münz, G., Li, S., & Carle, G. (2007, September). Traffic anomaly detection using k-means clustering. In *Gi/itg workshop mmbnet* (Vol. 7, No. 9).

1. Nella prima fase, si inizializzano in modo randomico k centri, dove k è arbitrario
2. Nella seconda fase, si associa a ciascun centro un cluster, e ogni punto viene assegnato al cluster corrispondente al centro più vicino
3. Nella terza fase, nel caso del kmeans, si ricalcolano i centri in base ai valori medi dei cluster, mentre, nel caso del kmedoids, diventa centro in ciascun cluster il punto, fra quelli osservati, che minimizza la distanza complessiva da tutti gli altri punti del cluster
4. La seconda e la terza fase si ripetono iterativamente, finché i centri non si stabilizzano, ovvero finché nessuna osservazione cambia cluster di appartenenza

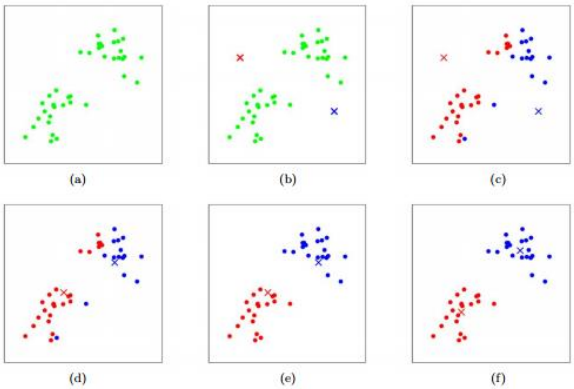


FIGURA 6 PROCESSO ITERATIVO DI POSIZIONAMENTO DEI CENTROIDI[60]

Ciascuno dei due metodi presenta dei vantaggi e degli svantaggi.
Il kmeans può calcolare dei centri più rappresentativi dei cluster ed è molto efficiente dal punto di vista computazionale, ma i centri di partenza sono posizionati in modo casuale, e pertanto i suoi risultati possono cambiare se viene eseguito più volte. Il kmedoids produce risultati stabili, ma non ha i vantaggi del kmeans. Inoltre, mentre, in presenza di outlier, il kmeans produce risultati distorti, il kmedoids è insensibile agli outlier. Sotto questo punto di vista la loro differenza è analoga a quella tra la media e la mediana. Utilizzeremo il kmedoids nel paragrafo 5.7 e nel paragrafo 6.4.

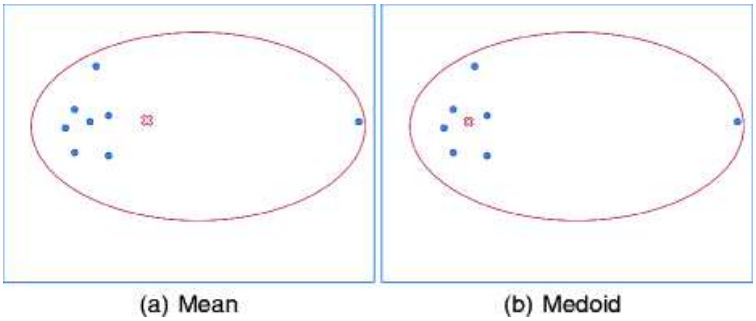


FIGURA 7 DIVERSA SENSIBILITÀ DEL KMEANS E DEL KMEDOIDS AGLI OUTLIER[61]

[60] Piech on kmeans (2013) https://stanford.edu/~cpiech/cs221/handouts/kmeans.html
[61] Kaufman, L., & Rousseeuw, P. J. (2009). *Finding groups in data: an introduction to cluster analysis*. John Wiley & Sons.

3.4 Valutazione del clustering

Esistono numerosi metodi per valutare la bontà del clustering, ovvero la qualità della ripartizione dei dati fra i gruppi. Molti di questi metodi tengono in considerazione due dimensioni:

- la coesione infra-cluster
- la separazione inter-cluster

L'indice comunemente più usato è quello della silhouette[62], compreso nel range [-1, 1], che consente di valutare il collocamento della singola osservazione, la qualità del singolo cluster e la bontà del clustering complessivo (nel qual caso viene denominata silhouette media).
Maggiore è la silhouette, migliore è la segmentazione dei dati.
La silhouette di una singola osservazione si definisce come segue:
a_i : distanza media tra l'osservazione i-esima e le osservazioni appartenenti al suo stesso cluster
b_i : distanza minima tra l'osservazione i-esima e le osservazioni esterne al suo cluster
s_i : indice di silhouette dell'osservazione i-esima

$$s_i = \frac{b_i - a_i}{\max(b_i, a_i)}$$

Il risultato della silhouette cambia in base alla metrica di distanza adottata per il calcolo di a e b.
La silhouette di un cluster si definisce come la media delle silhouette dei suoi elementi, mentre la silhouette del clustering complessivo, o silhouette media, è pari alla media delle silhouette di tutti i cluster ponderata per le loro dimensioni (ovvero alla media complessiva delle silhouette di tutte le osservazioni).
Esistono altri indici, come l'indice di Dunn, l'indice Calinski-Harabatsz e l'indice Davis-Bouldin.
Nell'utilizzo del clustering partizionale, si ricorre spesso alla within sum of squares (WSS) e alla between sum of squares (BSS), rispettivamente indicanti una misura della distanza tra i centri dei gruppi e i relativi elementi (inversamente proporzionale alla coesione interna dei cluster) e della distanza tra i centri e la media complessiva (direttamente proporzionale alla separazione fra i cluster). La loro somma si definisce come Total Sum of Squares (TSS).
Si vuole quindi massimizzare BSS e minimizzare WSS, ovvero si vuole portare

$$\frac{\text{BSS}}{\text{BSS} + \text{WSS}} = \frac{\text{BSS}}{\text{TSS}}$$

il più possibile vicino a 1.
Al crescere del numero di gruppi, sia la Silhouette sia la WSS si riducono.
In questa tesi, si adotta la WSS per individuare il numero ottimale di gruppi, utilizzando il clustering partizionale e il metodo Elbow, e la silhouette per valutare i risultati.

[62] Reynolds, A. P., Richards, G., de la Iglesia, B., & Rayward-Smith, V. J. (2006). Clustering rules: a comparison of partitioning and hierarchical clustering algorithms. *Journal of Mathematical Modelling and Algorithms*, *5*, 475-504. Thinsungnoena, T., Kaoungkub, N., Durongdumronchaib, P., Kerdprasopb, K., & Kerdprasopb, N. (2015). The clustering validity with silhouette and sum of squared errors. *learning*, *3*(7).

3.5 Il metodo Elbow

Sia nella famiglia degli algoritmi gerarchici sia in quella degli algoritmi partizionali è richiesto di specificare il numero di gruppi da costituire (l'altezza dell'albero o il numero di centri).
Sono stati proposti numerosi metodi per individuare il numero ottimo di cluster in una popolazione. Molto spesso, si assume che vi sia una conoscenza a priori del dominio del fenomeno studiato tale da possedere già questa informazione. In realtà, essendo il clustering uno strumento esplorativo, è piuttosto raro che l'analista sappia già in quanti gruppi dovrebbe dividersi la popolazione.
Molti studiosi tentano di individuare questo valore attraverso la Silhouette Media[63], ma tale metodo non sembra valido, data la sua persistenza a scegliere sempre k=2 (l'indice tende a diminuire sempre meno rapidamente al crescere di k, per k>2, ragion per cui quasi sempre la Silhouette media presenta un massimo globale in quel punto).
Spesso, utilizzando il kmeans, si fa riferimento al c.d. metodo del gomito (Elbow Method[64]). In base a questo metodo, si deve costruire un grafico che associa, per ogni numero di gruppi k, il corrispondente livello di WSS (Within Sum of Square), un indice di dispersione infragruppo.
Al crescere del numero di gruppi, infatti, la dispersione interna a ciascun gruppo si riduce, finché WSS = 0 per k = n.
Il grafico mostra così una curva decrescente, nel cui angolo più ripido si potrebbe visivamente individuare il punto di ottimo.
In linea generale, se si vuole calcolare analiticamente la coordinata del punto posizionato sull'angolo più ripido di una curva, è possibile calcolare la lunghezza di tutte le rette passanti per la curva e perpendicolari alla retta che collega i due estremi della curva, e poi individuare il punto della curva appartenente alla retta di lunghezza maggiore, ovvero che massimizza la distanza tra la curva e la retta degli estremi.

[63] Ishchenko, I., Globa, L. S., Buhaienko, Y., & Liashenko, A. (2019). Approach to determining the number of clusters in a data set.
Zhang, Y., Mańdziuk, J., Quek, C. H., & Goh, B. W. (2017). Curvature-based method for determining the number of clusters. *Information Sciences*, *415*, 414-428.
Yilmaz, S., Chambers, J., Cozza, S., & Patel, M. K. (2019, November). Exploratory study on clustering methods to identify electricity use patterns in building sector. In *Journal of Physics: Conference Series* (Vol. 1343, No. 1, p. 012044). IOP Publishing.
[64] Kodinariya, T. M., & Makwana, P. R. (2013). Review on determining number of Cluster in K-Means Clustering. *International Journal*, *1*(6), 90-95.
Shi, C., Wei, B., Wei, S., Wang, W., Liu, H., & Liu, J. (2021). A quantitative discriminant method of elbow point for the optimal number of clusters in clustering algorithm. *Eurasip Journal on Wireless Communications and Networking*, *2021*(1), 1-16.

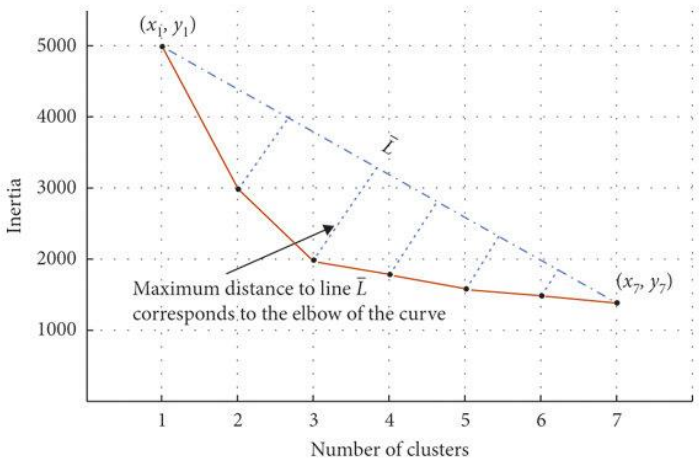


FIGURA 8 INDIVIDUAZIONE ANALITICA DEL PUNTO SULL'ANGOLO PIÙ RIPIDO DELLA CURVA[65]

Sempre con riferimento alla scelta del corretto numero di cluster, nella prassi si propone una semplice formula[66]:
k: numero di cluster
n: numero di individui

$$k \approx \sqrt{\frac{n}{2}}$$

Tale calcolo dovrebbe restituire un risultato simile a quello del metodo Elbow, se applicato sulla WSS dell'algoritmo kmeans.
Il metodo Elbow può essere adottato anche con riferimento al clustering gerarchico (utilizzando sulle ascisse il numero di cluster e sulle ordinate l'altezza del ramo in cui si taglia il dendrogramma) e nella scelta del numero di componenti della PCA (avendo sulle ascisse il numero di fattori e sulle ordinate la deviazione standard spiegata da ciascuna).
Con riferimento al metodo del gomito, si rimanda ai grafici nelle figure 21, 29, 40, 43 e 44.

[65] Shalaby, M., Belal, N. A., & Omar, Y. (2021). Data clustering improves Siamese neural networks classification of Parkinson's disease. *Complexity*, *2021*, 1-9.
[66] Madhulatha, T. S. (2012). An overview on clustering methods.

3.6 Il clustering DBSCAN e il rumore nei dati

Il DBSCAN clustering (Density-Based Spatial Clustering of Applications with Noise) appartiene alla famiglia del clustering basato sulla densità ed è stato introdotto da Ester et al.[67] nel 1996, come un metodo per l'analisi dei dati spaziali.
La sua principale applicazione consiste nella capacità di identificare sia gli outlier sia i noise points in aree a bassa densità, isolandoli o raccogliendoli in cluster di minor dimensione[68]. Questa peculiarità lo ha reso un algoritmo vastamente utilizzato.
Silver[69] evidenzia la differenza tra il rumore nei dati e gli outlier. Si tratta di due tipologie di osservazioni, che, in un campione, rendono distorta la percezione della reale distribuzione sottostante il fenomeno studiato. Tuttavia, mentre gli outlier rappresentano punti facilmente identificabili, in quanto aventi valori che fuoriescono dal range in cui si colloca la maggior parte dei dati, i noise points, ovvero il rumore, sono semplicemente valori rari che rendono distorta la distribuzione osservata senza discostarsi di molto dalla media.
In sintesi, gli outlier sono valori anomali, a volte informativi, che escono dall'intervallo dei dati, mentre i noise points sono valori inutili, che non rivelano nulla della distribuzione reale.
Con un metodo come quello dello Z-score è possibile identificare gli outlier, ma non i noise points. Il DBSCAN, invece, tenendo in considerazione la densità dei punti, è in grado di percepire la distorsione di entrambi.
L' algoritmo aggrega un punto a un cluster se sono rispettate due condizioni:

- La distanza tra il punto e il suo più vicino, appartenente al cluster, è sufficientemente bassa
- Il numero di punti da aggregare in quel cluster è sufficientemente alto

Come la maggior parte degli algoritmi di densità, il DBSCAN non richiede di specificare il numero di gruppi, che viene individuato in automatico.
Richiede tuttavia di specificare altri due parametri-soglia, inerenti alle due condizioni di cui sopra:

- La distanza minima fra due punti per essere considerati nello stesso cluster
- Il numero minimo di punti che deve essere compreso in un cluster per costituirlo

Questi due parametri vengono generalmente denominati Epsilon (indicante il raggio di un cerchio) e minPts (Minimun number of points).
Epsilon è la distanza minima, ovvero la soglia che consente di definire se due punti sono direttamente connessi (vicini) oppure no.
Nel DBSCAN, se un punto A ha una distanza dal punto B inferiore a Epsilon, e una distanza dal punto C superiore a Epsilon, ma la distanza tra B e C è inferiore a Epsilon, A e C sono considerati vicini, in quanto il punto B è vicino a entrambi. Si dice quindi che B è direttamente connesso ad A e a C, mentre A e C sono connessi fra loro per densità.

---

[67] Ester, M., Kriegel, H. P., Sander, J., & Xu, X. (1996, August). A density-based algorithm for discovering clusters in large spatial databases with noise. In *kdd* (Vol. 96, No. 34, pp. 226-231).
[68] Çelik, M., Dadaşer-Çelik, F., & Dokuz, A. Ş. (2011, June). Anomaly detection in temperature data using DBSCAN algorithm. In *2011 international symposium on innovations in intelligent systems and applications* (pp. 91-95). IEEE.
Çelik, M., Dadaşer-Çelik, F., & Dokuz, A. Ş. (2011, June). Anomaly detection in temperature data using DBSCAN algorithm. In *2011 international symposium on innovations in intelligent systems and applications* (pp. 91-95). IEEE.
[69] Silver, N. (2012). *The signal and the noise: Why so many predictions fail-but some don't*. Penguin.

Il DBSCAN classifica tutti i punti del dataset come di tre tipi:

- core points, ovvero direttamente connessi ad almeno minPts osservazioni nel dataset.
- density reachable points, ovvero direttamente connessi a più di zero osservazioni, di cui almeno una classificata come core point, ma a meno di minPts osservazioni. Possono considerarsi come punti di confine dei cluster.
- noise points, ovvero senza core points connessi, e per questo classificabili come outliers, in quanto non assegnati ad alcun cluster. Outlier e rumore sono rilevati nello stesso modo dal DBSCAN.

minPts può anche definirsi come il numero minimo di punti direttamente connessi che un punto deve avere per essere considerato un core point.
È sufficiente che vi sia un singolo core point per definire un cluster, ma tutti i core points reciprocamente connessi, direttamente o per densità, sono assegnati al medesimo cluster.

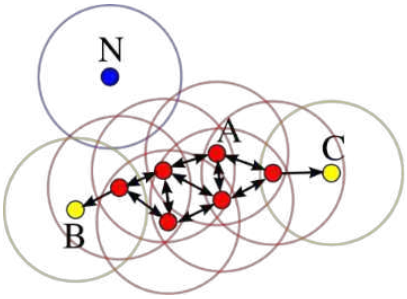


FIGURA 9 NELLA FIGURA, MINPTS=3, A È UN CORE POINT, B E C SONO DENSITY-REACHABLE POINTS MENTRE N È UN NOISE POINT[70]

In questa analisi adotteremo il DBSCAN nel paragrafo 5.6 sia per rimuovere le aziende sovra dimensionate rispetto alla media (gli outlier) sia per identificare eventuali aziende con caratteristiche eccessivamente diverse dalla maggioranza (il rumore).

[70] Toller, M. Anomalies in Data.

4. Metodologia adottata

L'oggetto della presente tesi è un metodo mediante il quale è possibile ottenere alcune informazioni utili per effettuare una buona analisi di settore. Tale metodo consiste in una procedura, che lascia ampi spazi di discrezionalità all'analista, ma allo stesso tempo consente di ottenere risultati obiettivi, se sussistono le premesse per poterlo adottare.

Il metodo in questione può sinteticamente suddividersi in sei fasi:

-nella prima, si selezionano le osservazioni inerenti al settore che si vuole analizzare, ovvero si costruisce il dataset sul quale si intende fare clustering. Tale dataset deve comprendere un numero sufficiente di osservazioni e deve contenere sia variabili di definizione strategica (che definiscono le scelte adottate dalla singola osservazione) sia variabili di valutazione delle performance (che possono essere messe a confronto per costruire dei benchmark tra le osservazioni)

-nella seconda fase, si osservano i dati, effettuando analisi delle distribuzioni delle variabili e delle loro relazioni (tra cui le analisi univariata e bivariata) per verificare se è opportuno adottare modelli di clustering, effettuare riduzioni dimensionali tramite l'analisi fattoriale, escludere alcune variabili o rimuovere delle osservazioni.

-nella terza fase, eventuale e non sempre necessaria, si applicano delle riduzioni dimensionali dei dati, con strumenti di analisi fattoriale, se il numero di variabili è eccessivo. In questa tesi, adottiamo la PCA per svolgere analisi fattoriale, ma sarebbe possibile ricorrere anche ad altri algoritmi, come la SVD e il t-SNE.

-nella quarta fase si verifica la presenza di outlier. Per individuarli all'interno di dataset di piccole o medie dimensioni si può adottare il clustering DBSCAN.

-nella quinta fase, si adotta il clustering partizionale per individuare i gruppi in cui dovrebbe idealmente suddividersi la popolazione.
Questa fase si suddivide a sua volta in tre momenti:

1. l'individuazione del numero ottimo di gruppi
2. la costruzione dei gruppi, e quindi l'individuazione dei centri
3. la valutazione della bontà del clustering, attraverso indici come la Silhouette

-nella sesta fase, si fa benchmarking dei cluster. Si recuperano i valori medi degli indici di performance di ciascun cluster, e, mediante una funzione di ranking, il Metodo Borda, si realizza una classifica dei cluster migliori, contenenti quindi valori di benchmark non solo per le variabili di valutazione di efficienza, ma anche per le variabili di definizione strategica. In questo modo, si ottengono informazioni sul comportamento medio dei soggetti che in media risultano più performanti.

Le premesse, ovvero le condizioni necessarie per poter svolgere con successo tale analisi, sono le seguenti:

- la presenza di indici di valutazione di strategia e di performance
- un sufficiente numero di osservazioni, e, di conseguenza, una conoscenza a priori il più possibile approfondita dell'attività di molte imprese già operanti nel settore
- una reale ripartizione della popolazione in cluster, verificabile tramite l'analisi della matrice delle distanze e gli indici di coesione dei cluster

Nei seguenti paragrafi verranno illustrati alcuni dei metodi statistici complementari al clustering, ma non direttamente collegati. In particolare, si vedranno lo Z-score, i coefficienti di correlazione, l'analisi fattoriale con PCA, e la funzione di ranking del metodo Borda.
Lo Z-score è utile per verificare la normalità di una singola variabile e per svolgere l'analisi fattoriale e di clustering.
I coefficienti di correlazione consentono di interpretare i risultati dell'analisi fattoriale.
La PCA, ovvero l'analisi fattoriale, consente di rendere più efficiente il funzionamento degli algoritmi di clustering, e necessita di una trasformazione dei dati tramite lo Z-score.
Il metodo Borda, infine, consente di svolgere un'analisi comparativa dei cluster, e di selezionare i benchmark di riferimento per il settore considerato. Può anche essere utilizzato per effettuare un ranking delle singole osservazioni, ma in questo modo non si è in grado di identificare dei benchmark realistici, cosa resa possibile dalla sua applicazione ai cluster.

Nei capitoli 4 e 5 vengono presentati due esempi di applicazioni del metodo appena descritto.
Si tratta di due dataset, uno sulle imprese operanti nella fabbricazione di motori e trasformatori elettrici, con 282 imprese, e uno sulle province italiane, con 86 osservazioni, entrambi ricavati da un dataset più grande, comprendente tutte le imprese operanti nella filiera di produzione di apparecchiature elettroniche (1254 osservazioni).
Il dataset di settore consentirà di individuare una buona strategia di investimento, per ogni segmento del mercato delle imprese con codice Ateco 271100.
Il dataset delle province consentirà, invece, di individuare le posizioni geografiche migliori in cui localizzare un'impresa entrante nel settore medesimo, a fronte di informazioni riferite alle diverse filiere produttive in cui il settore può inserirsi.

4.1 Lo Z-score

Prima di sottoporre una serie di variabili a un algoritmo di clustering[71] o di analisi fattoriale, risulta buona pratica standardizzare ciascuna variabile tramite lo Z-score[72], ovvero trasformare i suoi valori in valori appartenenti alla distribuzione Z (la Normale Standard di Gauss).
Lo Z-score consente infatti di svolgere tre operazioni importanti e di prassi:

- rendere comparabili le magnitudini di variabili diverse, in modo tale da non far prevalere il valore di quelle con una variabilità maggiore (non ha senso utilizzare insieme il fatturato di un'azienda e un valore percentuale per calcolare una metrica di distanza, senza prima aver normalizzato le due variabili) [73]
- verificare la normalità della distribuzione di ciascuna variabile (molti modelli, come gli algoritmi di clustering partizionale, si ritengono più validi se applicati su distribuzioni normali)
- individuare gli outlier di una variabile: secondo una prassi comune si considerano outlier quei valori che, dopo essere stati standardizzati, fuoriescono dal range [-3, 3][74] (nel quale, in una gaussiana ideale, dovrebbe rientrare oltre il 99% delle osservazioni)

Data una singola variabile misurata in un campione, la formula dello Z-score si definisce come segue:
x_i: osservazione i-esima della variabile *x*
μ: media campionaria della variabile *x*
σ: deviazione standard della variabile *x*
z_i: valore standardizzato di x_i

$$z_i = \frac{x_i - \mu}{\sigma}$$

Dopo aver standardizzato la variabile, possiamo verificare la normalità della sua distribuzione in quattro modi:

- osservando i quantili e la deviazione standard della variabile standardizzata: La mediana dovrebbe essere vicina alla media, ovvero nulla, il massimo e il minimo dovrebbero rientrare nel range [-3, 3], il primo e il terzo quantile dovrebbero risultare simili in modulo e la deviazione standard dovrebbe tendere a 1. Se pochi punti usciranno dal range [-3, 3], ma le altre condizioni saranno soddisfatte, si sarà in presenza di outlier.
- osservando il grafico della pdf (funzione di densità di probabilità) della variabile standardizzata: più la pdf assume la forma di una campana simmetrica, più la distribuzione è vicina a una distribuzione normale.
- osservando il grafico qq-plot della variabile standardizzata: se la maggior parte dei punti della variabile si posiziona sulla retta y=x, indicante i quantili di una distribuzione normale, la

[71] Mohamad, I. B., & Usman, D. (2013). Research Article Standardization and Its Effects on K-Means Clustering Algorithm. *Research Journal of Applied Sciences, Engineering and Technology*, *6*(17), 3299-3303.
[72] Chiaramonte, L., Croci, E., & Poli, F. (2015). Should we trust the Z-score? Evidence from the European Banking Industry. *Global Finance Journal*, *28*, 111-131.
[73] Skiena, S. S. (2017). *The data science design manual* Cap. 4.3: Z-scores and Normalization. Springer.
[74] Kannan, K. S., Manoj, K., & Arumugam, S. (2015). Labeling methods for identifying outliers. *International Journal of Statistics and Systems*, *10*(2), 231-238.

distribuzione della variabile sarà la stessa. Se pochi punti vi si discostano, gli stessi potranno qualificarsi come outlier.

- utilizzando il KS-test (Kolmogorov -Smirnov Test) avente come ipotesi nulla la non normalità della variabile. Se il p-value del test sarà inferiore al livello di significatività fissato, si potrà, per quel livello di significatività, rifiutare l'ipotesi nulla.

4.2 I coefficienti di correlazione

Ci serviamo del coefficiente di Pearson e del coefficiente di Spearman per valutare il grado di correlazione tra ogni coppia di variabili[75]. Entrambi rientrano nel range [-1, 1], dove gli estremi rappresentano massima correlazione (positiva e negativa), e 0 rappresenta correlazione nulla.
Il coefficiente di Pearson è una misura classica del livello di correlazione tra due variabili, ma misura la linearità di tale relazione. Il coefficiente di Spearman è anch'esso un indice di correlazione, ma a differenza del coefficiente di Pearson è in grado di cogliere relazioni non lineari[76]. Ciò significa che, in presenza di un coefficiente di Spearman notevolmente maggiore del coefficiente di Pearson, probabilmente ci si trova di fronte a una relazione non lineare, mentre, quando i due indici sono vicini, la relazione fra le variabili risulta lineare.
Date due variabili x, y e le rispettive deviazioni standard σ in un campione di numerosità n, i due coefficienti si definiscono come segue.
Definizione del coefficiente di Pearson:

$$r_p = \frac{cov_{xy}}{\sigma_x \sigma_y}$$

Dove:

$$cov_{xy} = \sum \frac{(x_i - \mu_x)(y_i - \mu_y)}{n-1}$$

Definizione del coefficiente di Spearman:

$$r_s = 1 - \frac{6 \sum d_i^2}{n(n^2-1)}$$

Dove:

$$d_i = rank(x_i) - rank(y_i)$$

I risultati dell'analisi bivariata dovrebbero riflettere quelli dell'analisi fattoriale[77].

[75] Skiena, S. S. (2017). *The data science design manual* Cap. 2.3: Correlation Analysis. Springer.
[76] Hauke, J., & Kossowski, T. (2011). Comparison of values of Pearson's and Spearman's correlation coefficients on the same sets of data. *Quaestiones geographicae*, *30*(2), 87-93.
[77] Wu, Y., Wang, Q., & Shi, Y. (2021). Research on Principal Component Feature Extraction Method Based on Improved Pearson Correlation Coefficient Analysis. In *Advances in Intelligent Information Hiding and Multimedia Signal Processing: Proceeding of the 16th International Conference on IIHMSP in conjunction with the 13th international conference on FITAT, November 5-7, 2020, Ho Chi Minh City, Vietnam, Volume 2* (pp. 82-87). Springer Singapore.

4.3 La Principal Component Analysis

L'analisi fattoriale è uno strumento di riduzione delle dimensioni dei dati, utile per rappresentare un certo numero di variabili c con un minor numero di fattori m senza perdere una quantità eccessiva di informazioni. Viene, spesso, adottata in via preliminare rispetto a un algoritmo di clustering, per ridurre il costo computazionale e ottenere un maggior livello di interpretabilità dei risultati.
In questo caso, adottiamo il metodo della PCA (Principal Component Analysis[78]) per calcolare le componenti principali presenti nei dati. Altri algoritmi di analisi fattoriale comunemente adottati sono la SVD e il t-SNE. Le componenti principali sono combinazioni lineari di tutte le variabili considerate, e ciascuna variabile contribuisce in una certa misura alla variabilità di ciascuna componente, ovvero ogni componente principale spiega una certa percentuale della variabilità complessiva.
Per fare in modo che ogni variabile incida nella stessa misura sulla variabilità complessiva dei dati, è necessario normalizzare tutte le variabili considerate attraverso il metodo dello Z-score.

Il grafico biplot[79] mostra ogni variabile come un vettore, ed evidenzia il grado di discostamento dalle direzioni delle altre variabili. Nel grafico, direzioni parallele (anche opposte) rappresentano grande correlazione tra due variabili, mentre direzioni perpendicolari rappresentano assoluta indipendenza tra le stesse. Al crescere del numero di variabili che vanno nella stessa direzione, cresce l'importanza della componente principale che le rappresenta.
A tal proposito si rimanda alle tabelle e ai biplot nelle figure 20, 22 e 39.
Le componenti principali (o fattori) sono ordinate in ordine decrescente in base alla percentuale di variabilità complessiva che spiegano (PC1, spiega una percentuale maggiore della variabilità complessiva spiegata da PC2, per PC2 vale lo stesso rispetto a PC3 e così via) e la variabilità spiegata da ciascuna componente principale non è spiegata da nessuna delle altre.
Il problema della scelta delle componenti principali si traduce quindi nel problema di scegliere quante usarne, in quanto ciascuna componente spiega una percentuale di variabilità minore della precedente e superiore alla successiva.
Un metodo per scegliere quante componenti principali usare può essere il metodo Elbow.

[78] XIE, X. (2019). Principal component analysis. *Wiley interdisciplinary reviews*.
[79] Frutos, E., Galindo, M. P., & Leiva, V. (2014). An interactive biplot implementation in R for modeling genotype-by-environment interaction. *Stochastic Environmental Research and Risk Assessment*, *28*, 1629-1641.

4.4 Il Metodo Borda per il benchmarking

Quando si svolgono analisi di benchmarking, spesso si deve tenere in considerazione una serie di punteggi, indipendenti gli uni dagli altri, per stilare una classifica delle migliori e delle peggiori performance complessive. Questo problema si pone non solo per il benchmarking degli individui nella popolazione, ma anche per il benchmarking dei diversi cluster in cui la popolazione è segmentata. Vi sono numerosi metodi per individuare dei benchmark in un settore[80].
È possibile utilizzare una funzione di ranking, come il Metodo Borda[81], per calcolare un punteggio univoco da associare a ciascun cluster.
Una funzione di ranking consente di ordinare un certo numero di giocatori per il punteggio che hanno ottenuto, mentre una funzione di scoring ha lo scopo di riassumere una serie di variabili in un unico valore. Il metodo Borda consente di assolvere entrambe le funzioni, a partire dal rango assegnato a una serie di soggetti. Il rango si definisce come la posizione che un soggetto occupa in un dato ordinamento[82]. In un ordinamento di n soggetti, il rango massimo è pari a n, ed è il peggiore, mentre il rango minimo è pari a 1, ed è il migliore.
Si distingue il Metodo Borda lineare dal Metodo Borda non lineare.
Il questo caso adottiamo il Metodo Borda lineare, che segue questo procedimento:
Dati n individui, a ciascuno dei quali è assegnata una serie di punteggi, per ogni punteggio, a ciascun individuo sono attribuiti (n - r) punti, dove r è il rango del dell'individuo, per quel punteggio, rispetto a tutti gli altri concorrenti.
Il punteggio complessivo di ciascun individuo è dato infine dalla somma dei punti derivanti dal rango di ciascun punteggio. In questo modo, è possibile calcolare uno score complessivo rappresentativo di tutti gli score considerati.
In altri termini, dato un numero di competizioni v e un numero di concorrenti n, il punteggio Borda S del soggetto c sarà dato da:

$$S_c = \sum_{i=1}^{v} n - r_i(c)$$

Dove $r_i(c)$ restituisce il rango del concorrente c nella competizione i.

[80] Sarkis, J., & Talluri, S. (2004). Performance based clustering for benchmarking of US airports. *Transportation Research Part A: Policy and Practice*, *38*(5), 329-346.
[81] Wu, W. W. (2011). Beyond Travel & Tourism competitiveness ranking using DEA, GST, ANN and Borda count. *Expert Systems with Applications*, *38*(10), 12974-12982.
[82] Treccani (2013) Enciclopedia della Matematica, voce "rango" in statistica

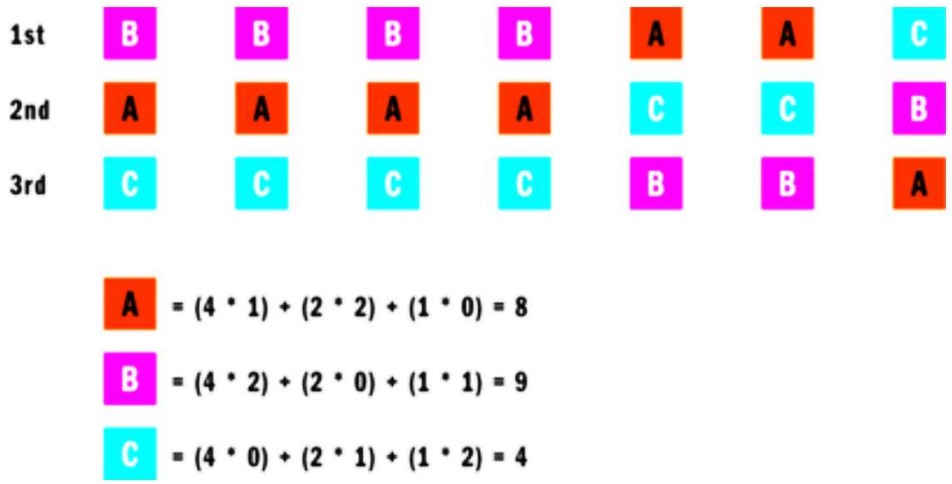


FIGURA 10 ESEMPIO DI CALCOLO DI TRE PUNTEGGI BORDA IN UNA COMPETIZIONE CON TRE CONCORRENTI E SETTE CONFRONTI[83]

Nel metodo Borda non lineare, invece, ogni punteggio (n - r) verrebbe moltiplicato per un peso derivante dal punteggio originale standardizzato, attraverso un altro metodo di normalizzazione, come ad esempio lo Z-score.
Tale metodo è più utile se si vuole evitare che due individui ricevano lo stesso punteggio.
Nelle analisi di settore, possiamo stilare la classifica delle imprese più performanti trattando ogni indice di bilancio come uno score rappresentativo delle performance dell'azienda in una particolare dimensione di valutazione[84] (es. solidità finanziaria, efficienza produttiva ecc.).
Con il Metodo Borda possiamo aggregare tutti gli score in uno solo, e possedere così un indice di valutazione generale della singola impresa. Nessuno score, tuttavia, ha valore se non vi è un ranking associato, e quindi se non vi è un ordinamento e un confronto tra diverse imprese.
Se vogliamo comprendere quale cluster di aziende è il migliore, dobbiamo affidarci ai valori medi delle performance dei cluster[85], e quindi ai loro centri.

[83] Kaner, C., Bach, J., & Pettichord, B. (2008). Test Run: Group Determination in Software Testing. MSDN Magazine
[84] Ecer, F., Böyükaslan, A., & Hashemkhani Zolfani, S. (2022). Evaluation of cryptocurrencies for investment decisions in the era of Industry 4.0: A borda count-based intuitionistic fuzzy set extensions EDAS-MAIRCA-MARCOS multi-criteria methodology. *Axioms*, *11*(8), 404.
[85] Dai, X., & Kuosmanen, T. (2014). Best-practice benchmarking using clustering methods: Application to energy regulation. *Omega*, *42*(1), 179-188.

5. Un primo esempio: quale strategia di investimento adottare?

Scopo di questo capitolo è di presentare una semplice analisi di settore, svolta secondo la procedura e gli strumenti descritti in precedenza. Intendiamo qui segmentare una popolazione di imprese in base alle loro strategie e performance di investimento, per identificare la combinazione di risorse mediamente più vicina a quella selezionata dalle imprese più performanti.

Il settore scelto è quello classificato attraverso il codice Ateco 271100: imprese operanti nella produzione di motori, generatori e trasformatori elettrici.

In base a un'ipotesi semplificatrice, la filiera cui assumiamo che il settore appartenga è quella di tutti i codici Ateco 27****.

Come si può intuire, entrambe le classificazioni sono approssimative e imperfette: vi sono imprese produttrici di motori elettrici che non sono impegnate nella produzione di trasformatori, e vi sono filiere produttive, cui le aziende della categoria 271100 potrebbero appartenere, comprendenti altre aziende non comprese nella macrocategoria 27****.

Tale approssimazione è necessaria, per tre ragioni:

1. In primo luogo, non si dispone nel caso specifico di una classificazione merceologica delle attività di queste imprese più dettagliata di quella dei codici Ateco.
2. In secondo luogo, le filiere produttive a cui le imprese che svolgono una data attività potrebbero appartenere sono molteplici, ed è prudente limitarsi a considerare solo i settori comprendenti la maggior parte dei fornitori e dei clienti tipici e diretti del settore studiato, ovvero i settori più simili, e per questo appartenenti alla stessa macrocategoria.
3. In terzo luogo, al crescere della precisione con cui si definisce l'attività svolta nel settore in questione, si riduce drasticamente il numero di imprese che vi opera, e, di conseguenza, la significatività dell'analisi. Si deve quindi individuare un equilibrio tra significatività statistica e precisione definitoria dei dati.

Nello studio delle variabili, ovvero nelle analisi univariata e bivariata, si esaminano tutte le osservazioni presenti nella filiera. Nelle parti successive si utilizza un dataset comprendenti voci e indici di bilancio di 282 imprese appartenenti al settore selezionato.

Utilizzeremo gli indici di bilancio come variabili di performance, da massimizzare, e le altre variabili come variabili di definizione strategica. Queste variabili consentono infatti di comprendere quanto le imprese migliori del settore, mediamente, investono in personale, R&D, materie prime, beni strumentali, e così via.

Le variabili scelte concentrano quindi l'attenzione sugli investimenti effettuati dall'azienda.

Ogni cluster rappresenterà una tipica combinazione strategica, e la classifica dei cluster più performanti del settore verrà eseguita attraverso il Metodo Borda.

5.1 Dati impiegati

Il database Aida, realizzato e distribuito da Bureau van Dijk S.p.A., comprende oltre un milione di bilanci di imprese italiane, nonché di informazioni aggiuntive per ciascuna impresa, come la provincia in cui la sede legale dell'azienda è collocata, il numero di dipendenti e il relativo codice Ateco e NACE, con riferimento agli anni dal 2013 al 2022. Ogni osservazione rappresenta un'impresa.
Le variabili selezionate e le imprese considerate, filtrate per settore, dimensione o localizzazione geografica, dipendono dal tipo di indagine che si vuole svolgere.

I dati utilizzati in questa analisi comprendono 282 osservazioni, se riferiti al settore Ateco osservato e all'anno 2021, comprendente le aziende operanti nella fabbricazione di generatori, motori e trasformatori elettrici, mentre ne comprendono 1254 se consideriamo l'intera filiera produttiva, ovvero l'insieme di tutte le imprese impegnate nella fabbricazione di apparecchiature elettriche, come cavi o impianti, aventi 27 come prime due cifre dei rispettivi codici Ateco.
Le variabili qui considerate sono distinte in tre gruppi, che possono distinguersi secondo un criterio funzionale. Tutte le voci di bilancio sono riferite all'anno e sono espresse in migliaia di euro.

Variabili di definizione qualitativa (categoriche):

- Codice Ateco: codice a sei cifre identificativo di una attività produttiva, ovvero di un settore
- Provincia: provincia nella quale si posiziona la sede principale dell'azienda

Variabili continue di definizione strategica:

- Ricavi: fatturato dell'azienda
- EBITDA: Earning Before Interest Tax Depreciation: indicatore di redditività, pari alla differenza tra il valore della produzione e i costi esterni e del personale
- Numero di dipendenti: numero di lavoratori con contratto di lavoro dipendente (non i consulenti, che rientrano nei servizi e nelle spese di outsourcing)
- Salari e stipendi: spese per il personale
- Costi delle materie prime: spese per i beni di consumo
- Costi dei servizi: spese per il lavoro prestato da lavoratori non dipendenti dell'azienda
- Costi di ricerca e pubblicità: comprendono le spese in ricerca e sviluppo e quelle pubblicitarie
- Costi di impianto e ampliamento: comprendono le spese che si sostengono in modo non ricorrente in alcuni caratteristici momenti del ciclo di vita della società, quali la fase pre-operativa (cosiddetti costi di start-up) o quella di accrescimento della capacità operativa.
- Costi sostenuti per il godimento di beni di terzi: pagati per l'utilizzo di beni di proprietà di terzi soggetti
- Valore complessivo delle immobilizzazioni materiali: valore economico degli immobili dell'azienda, come capannoni e impianti.
- Valore complessivo dei diritti di brevetto: valore economico dei brevetti posseduti dall'azienda

- Imposte complessivamente versate (correnti, differite, anticipate): possono anche assumere valore negativo, quando è sorto un credito d'imposta e l'impresa ha optato per il riporto a nuovo.

Variabili continue di valutazione delle performance (tutte espresse come valori percentuali):

- ROI: Return On Investment: pari al rapporto tra il Risultato Operativo e il Capitale investito Netto
- turnover ratio: pari al rapporto tra i ricavi netti di vendita e le attività totali nette. Misura il numero di volte che il capitale investito si rinnova per effetto del realizzo di nuovi ricavi. É una misura di efficienza.
- indice di liquidità: rappresenta il rapporto tra le attività a breve termine e le passività a breve termine.
- indice di indipendenza finanziaria: è pari al rapporto fra il capitale proprio e il totale delle passività.

5.2 Analisi Univariata

Si riportano di seguito le statistiche descrittive dei dati grezzi delle variabili continue utilizzate:

	variabile	minimo	primoQ	mediana	terzoQ	massimo	Nsample	media	deviazioneStd
[1,]	"Ricavi"	"0"	"940"	"4128"	"17140"	"1360345"	"1253"	"23294.5882"	"82570.2478"
[2,]	"SalariStipendi"	"0"	"155"	"544"	"2144"	"213557"	"1253"	"2595.5443"	"9622.4106"
[3,]	"TotImposte"	"-9036"	"5"	"30"	"230"	"46565"	"1253"	"383.9633"	"1998.7067"
[4,]	"Costi_ricerca_pubb"	"0"	"0"	"0"	"0"	"10560"	"1253"	"135.2674"	"687.9452"
[5,]	"Diritti_brevetto"	"0"	"0"	"0"	"9"	"153194"	"1253"	"233.4709"	"4406.0963"
[6,]	"Dipendenti"	"0"	"7"	"18"	"60"	"5239"	"1253"	"66.4278"	"206.6678"
[7,]	"ROI"	"-28.98"	"3.5"	"8.57"	"15.38"	"29.66"	"1253"	"8.9633"	"9.9059"
[8,]	"EBITDA"	"-56918"	"54"	"272"	"1544"	"272033"	"1253"	"2249.988"	"10971.8002"
[9,]	"CostiMateriePrime"	"0"	"376"	"2063"	"9168"	"1196707"	"1253"	"14982.4374"	"59874.3862"
[10,]	"CostiServizi"	"2"	"196"	"751"	"2706"	"386508"	"1253"	"4036.6464"	"15475.3144"
[11,]	"CostiGodimentoBeniDiTerzi"	"0"	"19"	"66"	"245"	"24733"	"1253"	"324.0942"	"1309.017"
[12,]	"CostiImpiantoAmpl"	"0"	"0"	"0"	"0"	"3247"	"1253"	"10.099"	"113.5311"
[13,]	"ImmoMateriali"	"0"	"84"	"564"	"2787"	"372337"	"1253"	"4584.0623"	"18441.1519"
[14,]	"TurnoverRatio"	"0"	"0.65"	"0.92"	"1.24"	"3.86"	"1253"	"0.9642"	"0.4811"
[15,]	"IndLiquidit"	"0.02"	"0.87"	"1.26"	"2.02"	"9.42"	"1253"	"1.6901"	"1.3617"
[16,]	"IndiceIndipFinanz"	"-35.22"	"19.25"	"35.81"	"54.77"	"99.28"	"1253"	"37.7964"	"22.9666"

FIGURA 11 STATISTICHE DESCRITTIVE DEI DATI GREZZI DI TUTTE LE IMPRESE A LIVELLO DI FILIERA

Con riferimento al dataset impiegato, avente come righe 1253 imprese con codice Ateco 27****, si costruiscono i grafici delle pdf (probability density function) e dei qqplot (quantile-quantile plot) delle 16 variabili continue, prima senza applicare la trasformazione logaritmica e senza rimuovere gli outlier, e poi applicando entrambe le operazioni.

Si riportano di seguito le pdf normali e i qqplot dei dati grezzi:

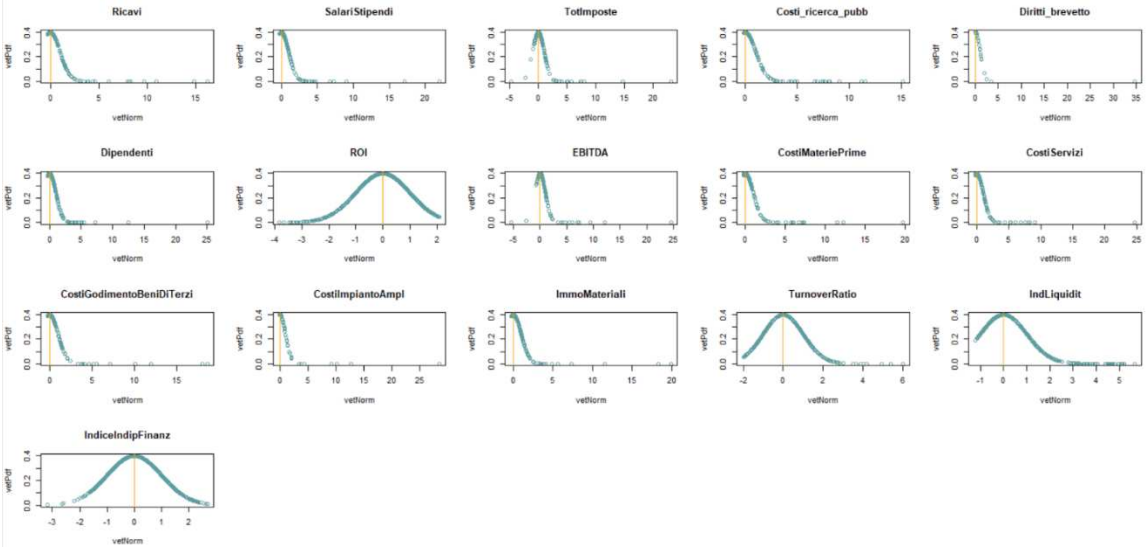


FIGURA 12 PDF NORMALI DELLE VARIABILI NEI DATI GREZZI SULLE IMPRESE DELLA FILIERA

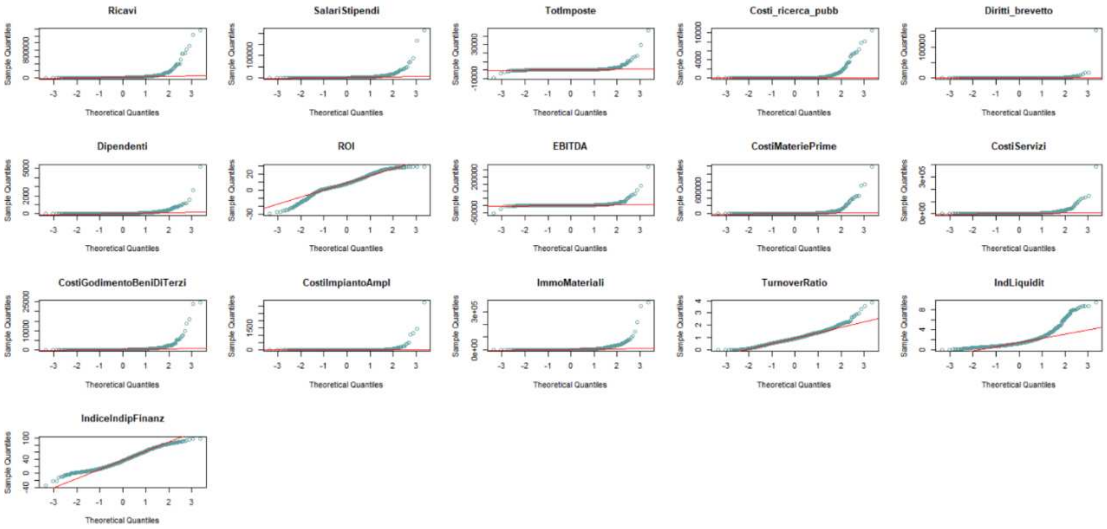


FIGURA 13 QQPLOT NORMALI DEI DATI GREZZI SULLE IMPRESE DI FILIERA

La maggior parte delle distribuzioni si discosta molto dalla forma della Normale.
Di frequente molte distribuzioni non si presentano come normali, ma come log-normali[86]. Ciò significa che, spesso, se si applica la trasformazione logaritmica a una variabile non distribuita normalmente, la variabile trasformata lo sarà, e, se ciò accade, è anche possibile assumere che la distribuzione della variabile non trasformata, a livello di popolazione, sia normale.

Inoltre, la presenza di outlier potrebbe contribuire a rendere distorta una distribuzione realmente normale.
In base al metodo dello Z score, si considerano outlier tutti i valori che, standardizzati in base alla media e alla deviazione standard della variabile, fuoriescono dal range [-3, 3]. Se rimuoviamo tali punti, vediamo che la distribuzione si avvicina maggiormente a una gaussiana ideale.

Si riportano di seguito i risultati sulle pdf e sui grafici qq:

[86] West, R. M. (2022). Best practice in statistics: The use of log transformation. *Annals of Clinical Biochemistry*, *59*(3), 162-165.

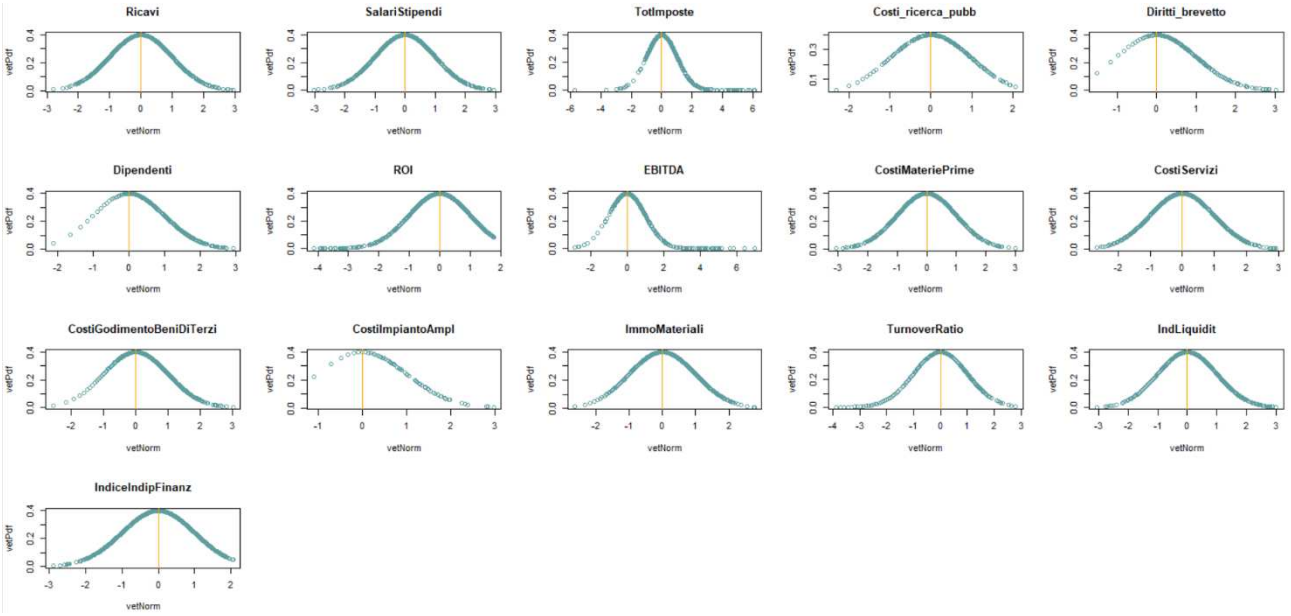


FIGURA 14 PDF NORMALI DELLE IMPRESE DI FILIERA, DOPO TRASFORMAZIONE LOGARITMICA E RIMOZIONE DEGLI OUTLIER

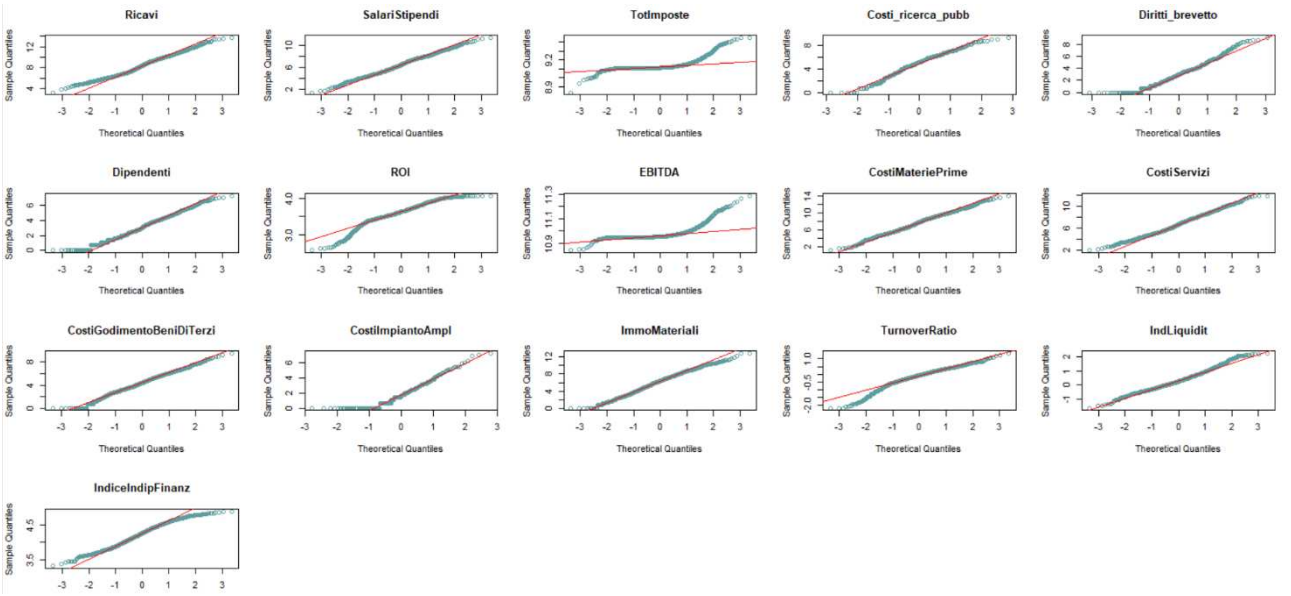


FIGURA 15 QQ-PLOT NORMALI DELLE IMPRESE DI FILIERA, DOPO TRASFORMAZIONE LOGARITMICA E RIMOZIONE DEGLI OUTLIER

A seguito della trasformazione operata e della rimozione degli outlier, la maggior parte delle variabili considerate si presenta come distribuita normalmente, per cui è possibile utilizzarle per effettuare analisi di clustering partizionale.

5.3 Analisi Bivariata

Si riportano qui le heatmap e i grafi dei livelli di correlazione, calcolati con Spearman e Pearson, tra ogni coppia di variabili standardizzate. Le correlazioni forti sono rappresentate nelle heatmap da colori scuri, e nei grafi da archi larghi e corti. Nelle heatmap, colori caldi indicano correlazioni positive, mentre colori freddi indicano correlazioni negative.

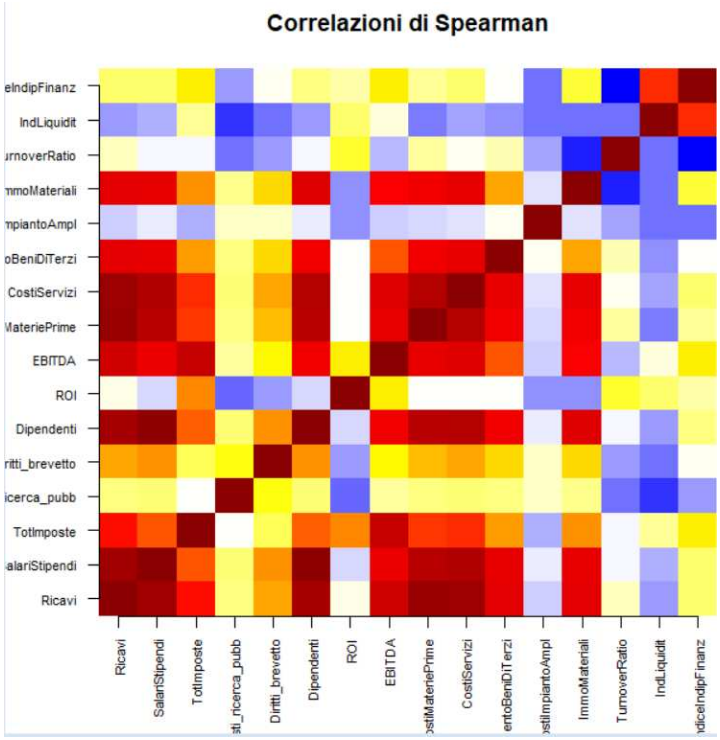


FIGURA 16 COEFFICIENTI DI SPEARMAN: COLORI SCURI IMPLICANO ALTA CORRELAZIONE

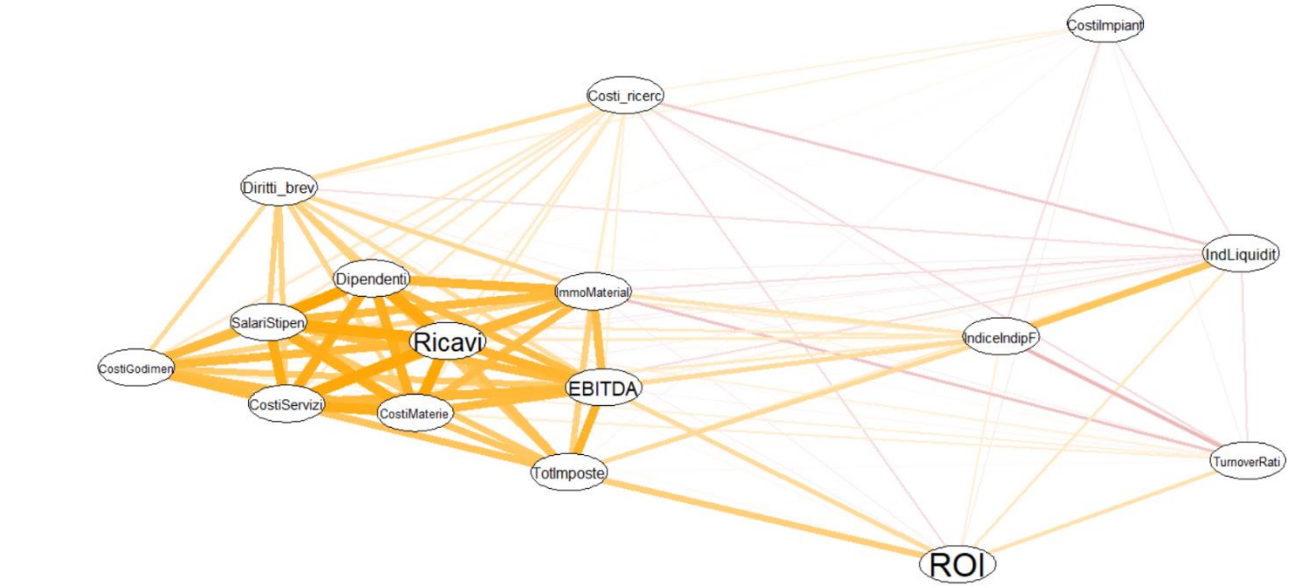


FIGURA 17 COEFFICIENTI DI SPEARMAN: ARCHI CORTI E SPESSI INDICANO FORTI CORRELAZIONI (POSITIVE E NEGATIVE)

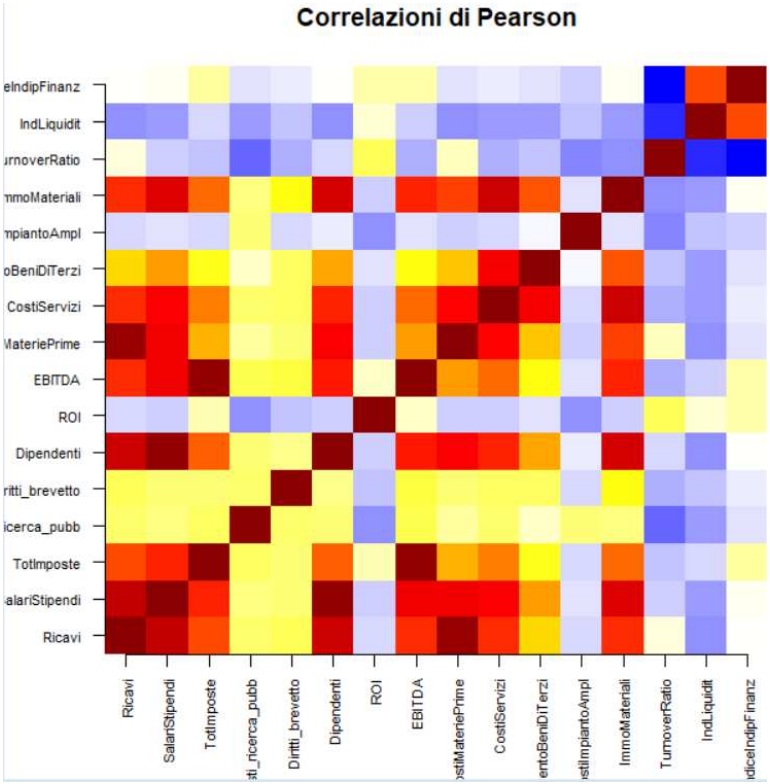


FIGURA 18 COEFFICIENTI DI PEARSON: COLORI SCURI INDICANO FORTI CORRELAZIONI

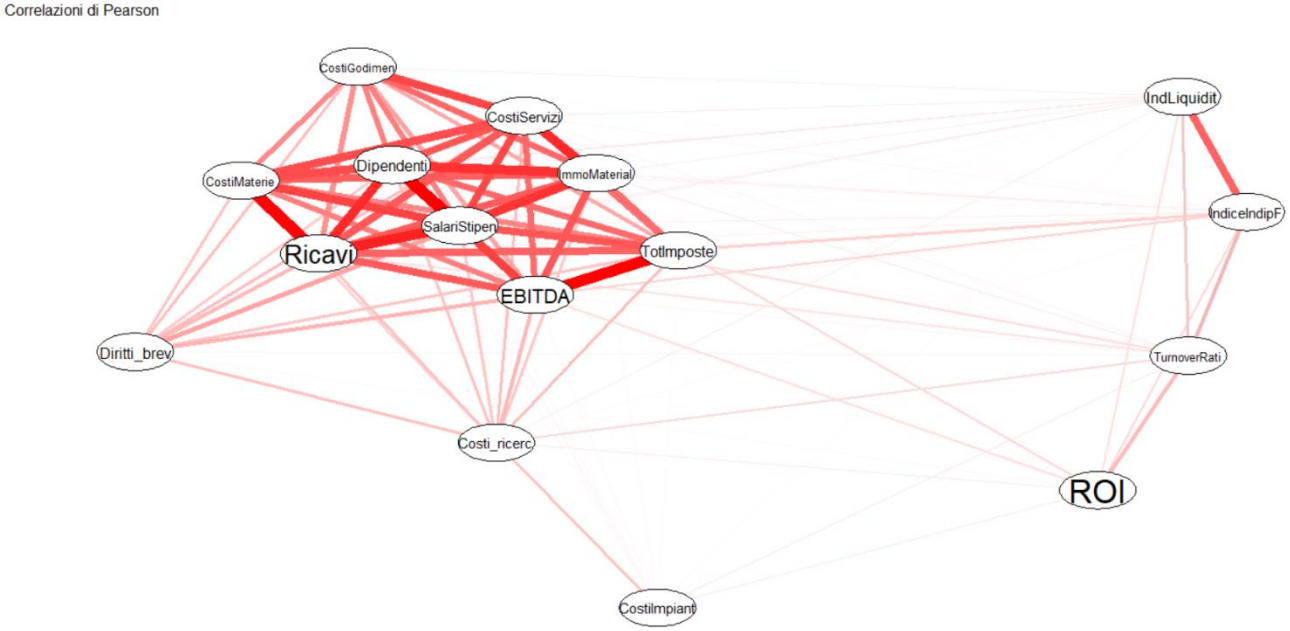


FIGURA 19 COEFFICIENTI DI PEARSON: ARCHI CORTI E SPESSI INDICANO FORTI CORRELAZIONI (POSITIVE E NEGATIVE)

5.4 Analisi Fattoriale

La PCA consente di effettuare una riduzione dimensionale del dataset.
Le 16 variabili possono assumere tre tipi di valori. Ci sono le variabili di performance, che sono espresse in termini percentuali, e che raramente escono dall'intervallo [-1,1]. Ci sono le variabili espresse in termini monetari, delle quali alcune possono assumere valori negativi (come le imposte pagate, che possono produrre la riscossione di un credito d'imposta) mentre altre hanno 0 come minimo (come il fatturato). Infine, c'è il numero di dipendenti, sempre maggiore di 1 ma spesso più contenuto delle altre variabili monetarie.
Il dataset sul quale viene applicata la PCA viene prima standardizzato sui valori della distribuzione Z, in modo tale da rendere confrontabili le loro varianze.
Notiamo che i risultati del biplot riflettono quelli dell'analisi bivariata:

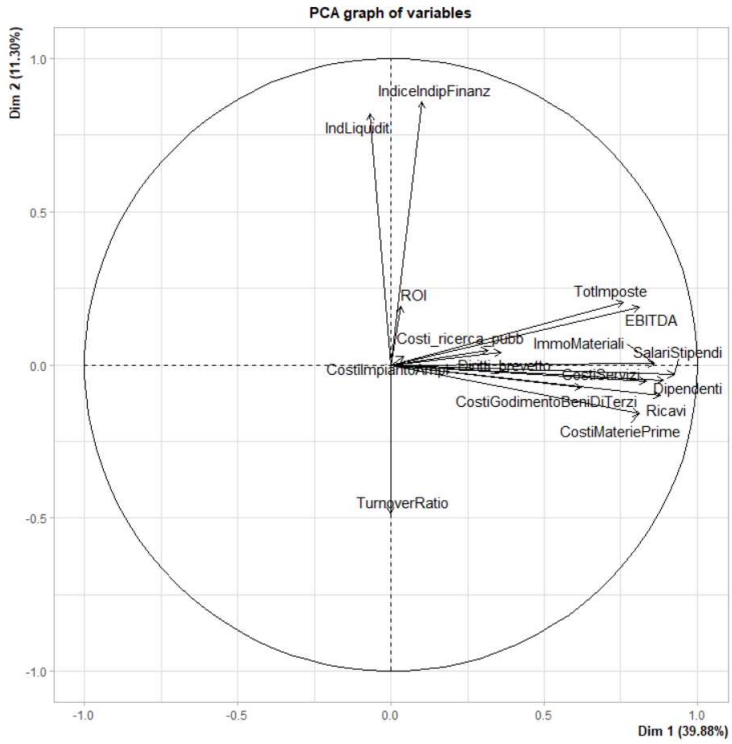


FIGURA 20 BIPLOT: VARIABILI CON DIREZIONI PARALLELE SONO RECIPROCAMENTE CORRELATE, VARIABILI CON DIREZIONI PERPENDICOLARI PRESENTANO CORRELAZIONE NULLA

Dal grafico, si può presumere che esistano due direzioni prevalenti, per cui sarebbe già possibile presumere che vi siano due componenti principali (le prime due, per la precisione) che spiegano in buona parte la variabilità complessiva.
È da notare come questa distinzione rifletta la separazione fatta a priori tra variabili indicative delle performance, misurate in termini percentuali, e variabili misurate in termini monetari, fatta eccezione per il numero di dipendenti, e comunque mai rappresentative di valori frazionari.
Il Turnover Ratio presenta una forte correlazione negativa con gli altri indici di bilancio, e quindi una direzione parallela anche se opposta a queste.
Il gruppo degli indici percentuali si mostra meno numeroso ma più compatto del gruppo delle voci di bilancio.

Per individuare il numero di componenti ottimale, ci si può affidare al metodo del gomito, utilizzando la curva rappresentante la relazione inversa tra deviazione standard complessiva spiegata da ciascuna componente e il numero di componenti da raccogliere.

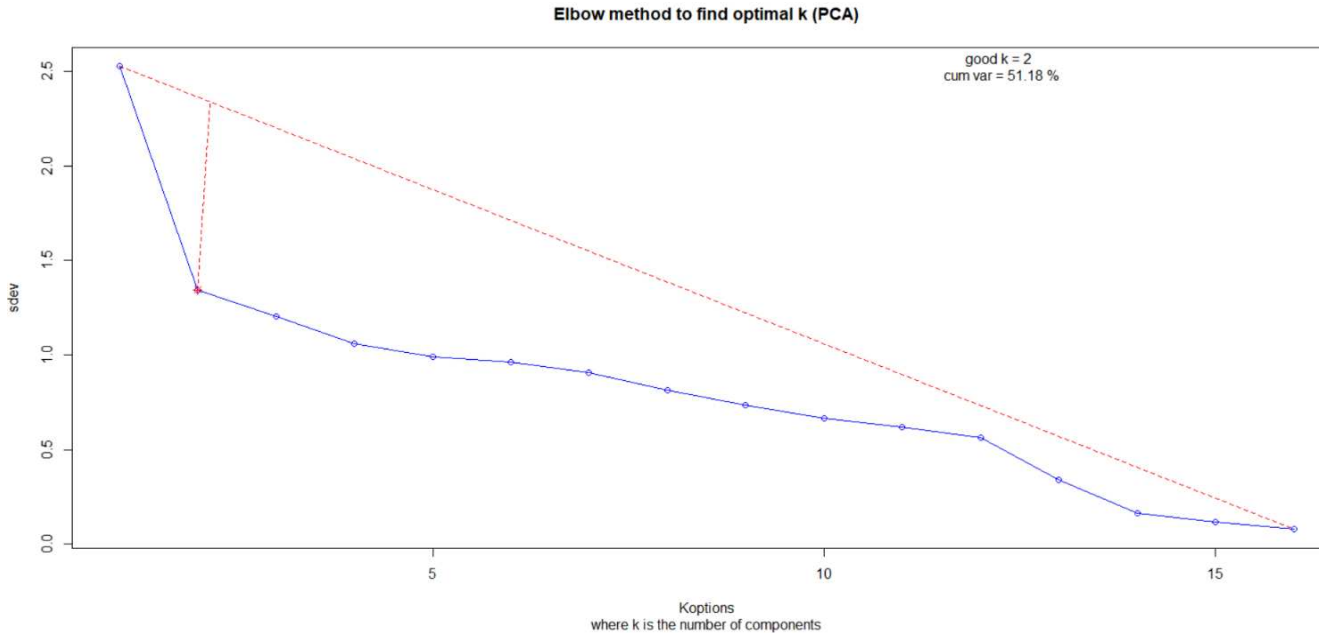


FIGURA 21 METODO ELBOW APPLICATO ALLO SCREEPLOT DELLE COMPONENTI PRINCIPALI (NOTA: LA RETTA CHE COLPISCE L'ANGOLO PIÙ RIPIDO DELLA CURVA NON SEMBRA PERPENDICOLARE IN QUANTO LE SCALE DEGLI ASSI DIFFERISCONO)

```
> summary(pca)
Importance of components:
                          PC1    PC2     PC3     PC4     PC5     PC6     PC7     PC8
Standard deviation     2.5259 1.3448 1.20295 1.05970 0.98915 0.96434 0.90815 0.81452
Proportion of Variance 0.3988 0.1130 0.09044 0.07019 0.06115 0.05812 0.05155 0.04146
Cumulative Proportion  0.3988 0.5118 0.60224 0.67242 0.73357 0.79169 0.84324 0.88471

                           PC9    PC10    PC11    PC12    PC13    PC14    PC15    PC16
Standard deviation     0.73662 0.66471 0.61800 0.56219 0.34064 0.16176 0.11870 0.07744
Proportion of Variance 0.03391 0.02762 0.02387 0.01975 0.00725 0.00164 0.00088 0.00037
Cumulative Proportion  0.91862 0.94623 0.97010 0.98986 0.99711 0.99874 0.99963 1.00000
```

FIGURA 22 DEVIAZIONE STANDARD, PERCENTUALE DELLA VARIANZA COMPLESSIVA E PERCENTUALE DELLA VARIANZA CUMULATA SPIEGATE DA CIASCUNA COMPONENTE PRINCIPALE

In questo caso le componenti principali che utilizzeremo saranno PC1 e PC2, che spiegano rispettivamente il 39.88% e il 11.30% della varianza complessiva.
Rinunciamo a una proporzione di informazione, al fine di guadagnare in termini di interpretabilità dei risultati e costo computazionale.

5.5 Costruzione della matrice delle distanze

Nei paragrafi precedenti si sono eseguite le analisi delle variabili sul dataset contenente tutte le imprese della filiera produttiva.
L'analisi di settore deve però svolgersi sul sotto-dataset contenente le imprese con codice Ateco 271100, ovvero impegnate nella fabbricazione di generatori, motori e trasformatori elettrici, trattandosi del settore che vogliamo analizzare.
Visivamente, possiamo intuire se vi sia una certa ripartizione del campione in più cluster, o se vi siano outlier fra i punti, osservando la matrice:

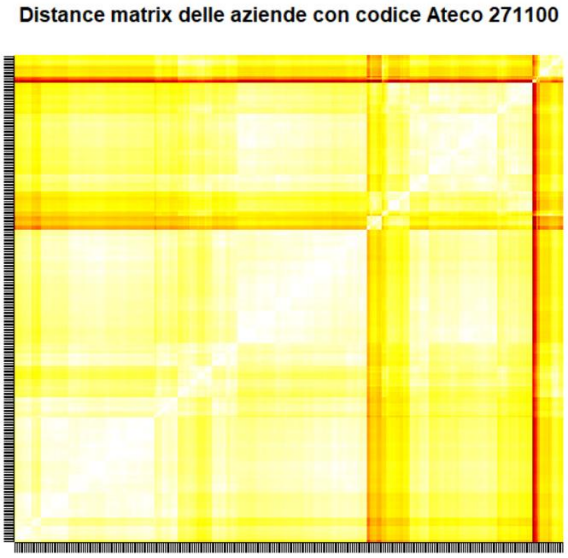


FIGURA 23 MATRICE DELLE DISTANZE CON DISTANZA EUCLIDEA CON RIGHE ORDINATE ATTRAVERSO IL CLUSTERING GERARCHICO CON COMPLETE LINKAGE: COLORI CHIARI IMPLICANO FORTE SOMIGLIANZA TRA LE OSSERVAZIONI SULLE RIGHE E SULLE COLONNE

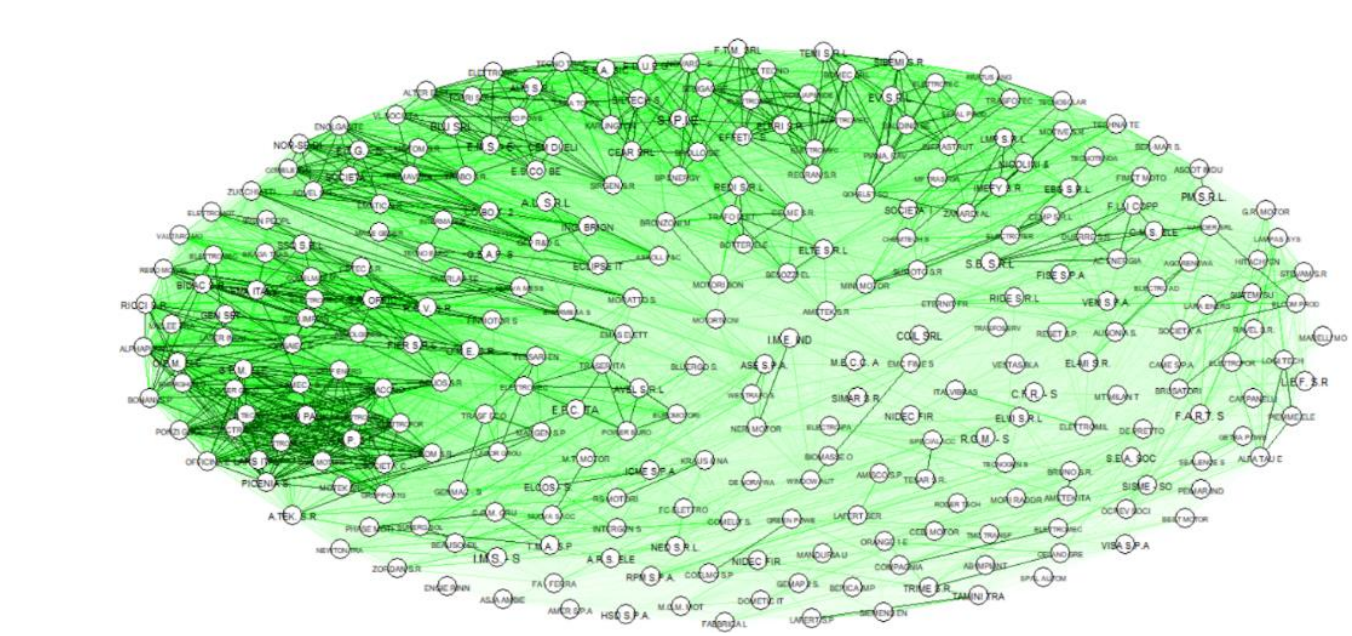


FIGURA 24 MATRICE DELLE DISTANZE CON DISTANZA EUCLIDEA SOTTO FORMA DI GRAFO: ARCHI CORTI E SCURI IMPLICANO FORTE SOMIGLIANZA TRA OGNI COPPIA DI OSSERVAZIONI

Dalla matrice delle distanze, vediamo che alcune osservazioni (nella heatmap le righe rosse, nel grafo i nodi senza nessun arco) sono molto lontane da tutte le altre. Conviene quindi provare a rimuovere degli outlier.
Inoltre, non vi è una sola area in cui i punti si concentrano (nella heatmap i quadrati bianchi sulla diagonale, nel grafo i gruppi di nodi di colore più scuro), perciò è probabile che esistano più sottogruppi nella popolazione campionaria. Vi sono probabilmente imprese di piccole, medie e grandi dimensioni in questo settore, ma vi potrebbero anche essere differenze ambientali, strutturali e strategiche, tali da motivare scelte di investimento completamente differenti.
I cluster consentiranno di definire le caratteristiche rilevanti di ciascun gruppo.

5.6 Rimozione outlier con DBSCAN

Prima di procedere alla successiva fase di segmentazione del campione in cluster, è necessario rimuovere gli elementi outlier e di rumore presenti nel dataset.
Si tratta di un dataset con sole 282 imprese, alcune delle quali potrebbero risultare outlier, in quanto molto più grandi delle aziende che sono appena entrate nel mercato o che presentano un comune volume di operazioni. Altre imprese potrebbero invece risultare significativamente diverse dalla maggior parte di quelle operanti nel settore, pur senza uscire dal range dei valori mediamente osservati, e costituire fonte di rumore.
È pertanto necessario identificare entrambe le cause di distorsione della distribuzione campionaria.
Un metodo di clustering utile a tal fine potrebbe essere il DBSCAN (Density-Based Spatial Clustering of Applications with Noise).
A differenza degli algoritmi di clustering più comuni, nei quali il numero di cluster deve essere specificato dall'analista, il DBSCAN individua automaticamente quel numero, e la sua peculiarità consiste nella valorizzazione della nozione di densità.

Prima di eseguire il clustering, conviene scalare i dati (le colonne PC1 e PC2) con lo Z-score, in modo tale da normalizzare l'impatto dei due fattori sul calcolo della distanza.
Fissiamo eps=1 (una distanza minima pari alla deviazione standard della Normale) e minPts=1 (per isolare i noise points). Otteniamo il seguente risultato:

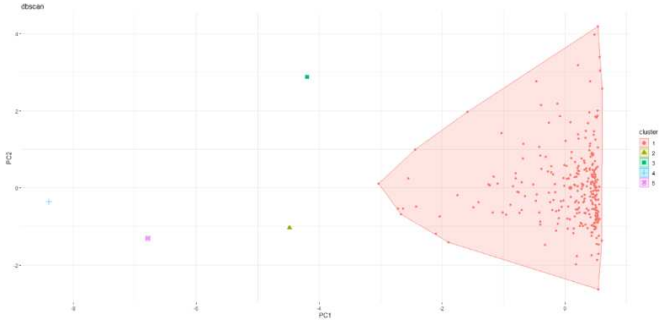


FIGURA 25 RISULTATO DEL CLUSTERING DBSCAN: ISOLAMENTO DEI NOISE POINTS

Nel grafico non sembrano esserci punti di rumore, in quanto i punti isolati si posizionano tutti al di fuori dell'area in cui si colloca la maggior parte delle imprese.
Fissando minPts=2 otteniamo un risultato analogo:

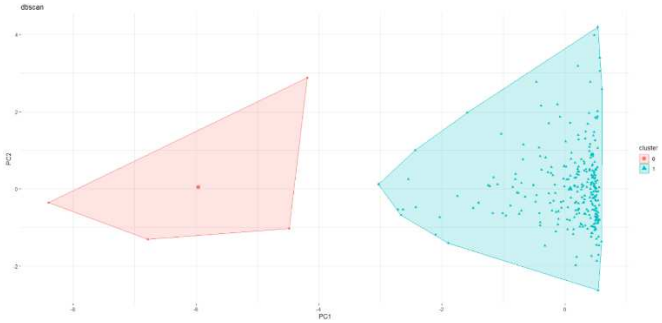


FIGURA 26 RISULTATO DEL CLUSTERING DBSCAN: RACCOLTA DEI NOISE POINTS IN UN UNICO CLUSTER

Per cui rimuoviamo 4 outlier, mantenendo un dataset di 278 osservazioni.
La silhouette media è dell'80%.

5.7 Cluster di imprese

Vogliamo individuare i cluster effettivamente presenti nella popolazione.
Questa fase, rappresenta quella di segmentazione effettiva dei gruppi di imprese, e costituisce l'elemento centrale dell'analisi. Tramite questi algoritmi potremo:

- individuare il numero ottimo di cluster
- assegnare ogni osservazione al cluster che meglio la rappresenta
- calcolare i valori medi di ciascuna variabile per ciascun gruppo

Possiamo ricorrere ai grafici della silhouette media, calcolata per ogni livello di k, per confrontare il clustering gerarchico (con euclidean distance e complete linkage) con il kmeans e il kmedoids. Vediamo che le performance dei tre algoritmi risultano molto simili.

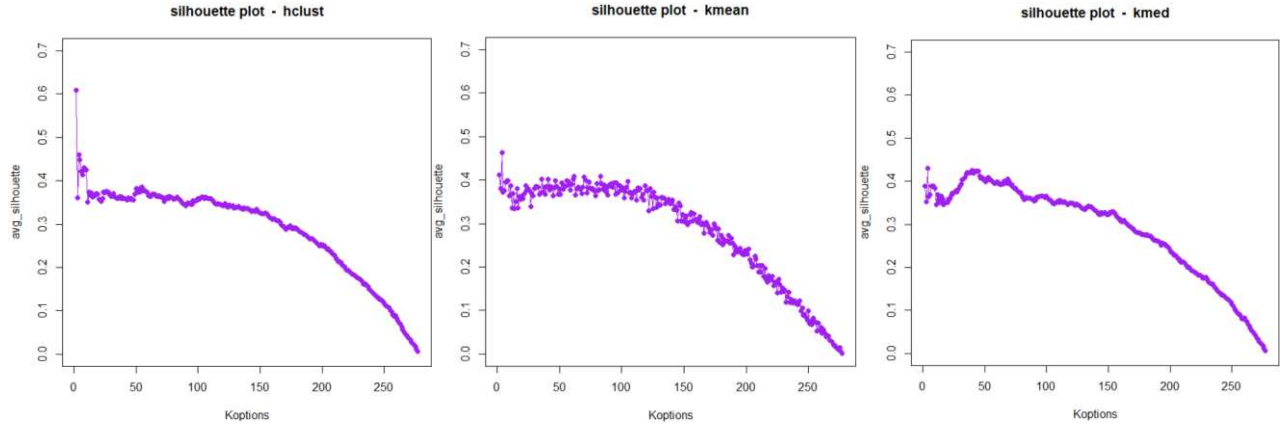


FIGURA 27 LIVELLI DI SILHOUETTE MEDIA PER OGNI NUMERO DI CLUSTER CON CLUSTERING GERARCHICO, KMEANS E KMEDOIDS

Adotteremo il kmeans e il metodo del gomito per individuare il numero ottimo di gruppi, e utilizzeremo il kmedoids per formare i gruppi finali.
Con il metodo del gomito, possiamo minimizzare contemporaneamente la dispersione interna dei cluster WSS e il numero di gruppi k.

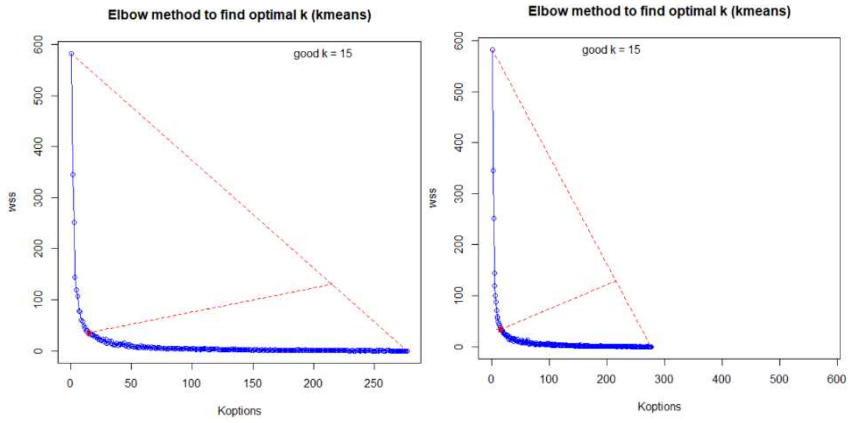


FIGURA 28 RISULTATO DEL METODO ELBOW PER L'INDIVIDUAZIONE DEL NUMERO OTTIMO DI GRUPPI CON WSS E KMEANS (NEL GRAFICO A SINISTRA LE SCALE DEGLI ASSI DIFFERISCONO)

Il risultato del metodo elbow è 15.
Notiamo che si ricava un risultato simile utilizzando la formula

$$k \approx \sqrt{\frac{n}{2}} = \sqrt{\frac{278}{2}} = 11.79$$

Il risultato del gomito sembra quindi verosimile.

Una volta scelto il valore di k, non resta che trovare i centri dei k cluster con il kmedoids, e poi i centroidi dei cluster con il kmeans, dopo averli inizializzati come i centri del kmedoids.

Ecco i cluster costituiti:

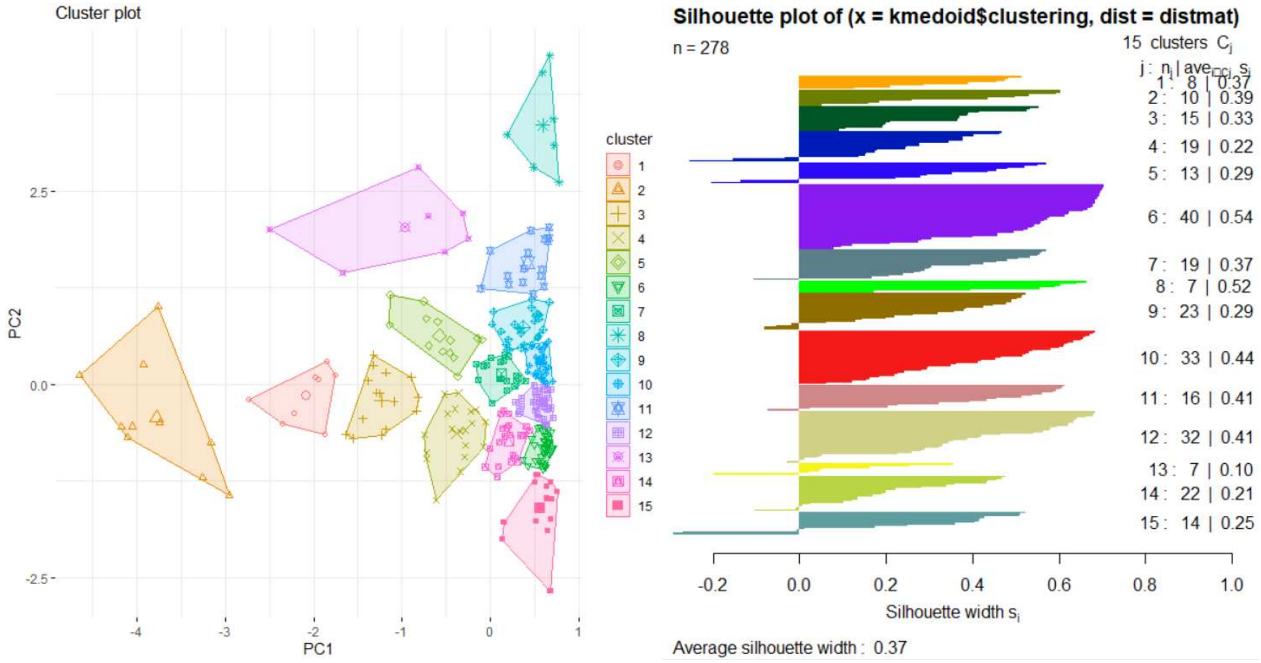


FIGURA 29 RISULTATI DEL CLUSTERING KMEDOIDS E SILHOUETTE MEDIA DI CIASCUN CLUSTER

La silhouette media è del 37%. Lo stesso indice consente di misurare la qualità di ciascun gruppo.

La numerosità dei cluster è ripartita in modo abbastanza equo.

Possiamo osservare i valori medi delle variabili strategiche di ciascun cluster per mezzo del suo centro:

	Group.1	Ricavi	SalariStipendi	TotImposte	Costi_ricerca_pubb	Diritti_brevetto	Dipendenti	EBITDA
1	1	56118.12	2706.875	75.3750	1.50000	80.12500	77.12500	486.7500
2	2	38673.90	5757.500	1229.1000	35.70000	517.70000	152.20000	6494.6000
3	3	28524.07	2630.000	448.2667	104.73333	29.26667	73.20000	3276.4667
4	4	28072.53	3647.684	192.8421	77.57895	336.57895	91.94737	3321.3684
5	5	131033.23	20324.154	1074.3077	43.92308	185.38462	505.61538	9371.8462
6	6	19633.08	2459.125	447.5000	37.25000	19.65000	74.75000	2617.6500
7	7	74781.32	4639.789	1340.7895	417.10526	229.89474	110.05263	6508.6842
8	8	33453.86	2272.714	162.8571	15.85714	10.85714	60.57143	2959.4286
9	9	29358.48	2135.522	-154.2174	51.69565	123.69565	59.21739	-982.7826
10	10	20669.76	2516.879	335.6061	143.57576	47.30303	69.21212	2200.4545
11	11	26846.06	3606.312	236.1875	150.68750	13.68750	98.06250	1954.3750
12	12	20786.09	2382.000	417.9375	56.87500	145.96875	67.06250	2357.7500
13	13	24933.57	3473.857	445.8571	0.00000	54.14286	102.42857	3302.2857
14	14	29649.05	3571.727	331.3182	65.09091	279.81818	85.77273	2093.3182
15	15	34199.57	4856.000	1293.7857	1270.64286	524.21429	113.14286	7785.5000

	CostiMateriePrime	CostiServizi	CostiGodimentoBeniDiTerzi	CostiImpiantoAmpl	ImmoMateriali
1	49978.62	5001.000	437.12500	0.1250000	1955.750
2	20397.70	5391.100	292.40000	39.6000000	9373.000
3	17368.13	5166.333	440.20000	2.8666667	3380.533
4	15678.37	5261.000	275.84211	0.0000000	7311.211
5	81212.69	15988.308	963.23077	0.6923077	38339.846
6	11538.02	3478.800	365.05000	47.2750000	4244.900
7	55831.11	6992.211	744.21053	7.7894737	10265.368
8	26023.43	2738.571	28.42857	0.0000000	7148.286
9	21178.70	4105.652	235.13043	0.2173913	5239.783
10	12663.21	3372.000	209.00000	10.0303030	3936.788
11	14568.56	6719.062	230.68750	0.5625000	6075.750
12	12338.03	3753.781	327.81250	0.8125000	4552.469
13	13755.86	4234.143	311.42857	0.0000000	5526.429
14	18434.68	5331.455	450.50000	3.0454545	4800.864
15	12277.71	5932.500	708.35714	55.7142857	6414.429

FIGURA 30 VALORI MEDI NEI CLUSTER DELLE VARIABILI STRATEGICHE

Le rispettive variabili di performance risultano:

```
> centri[,indici]
         ROI TurnoverRatio IndLiquidit IndiceIndipFinanz
1   8.192500     0.9300000    1.786250          39.21125
2  13.998000     0.9580000    1.309000          39.63800
3   9.276000     1.1366667    1.576667          40.75133
4  10.526842     1.1257895    1.496316          49.74579
5   8.817692     0.9692308    1.450000          48.65692
6  10.280250     1.0177500    1.634000          41.37300
7   9.700526     0.9626316    1.965789          43.22421
8  11.700000     1.2528571    1.597143          35.85000
9   7.111304     0.9043478    1.684348          42.31609
10  8.684545     1.0560606    1.410303          37.72152
11 10.242500     0.9943750    1.549375          36.22125
12  8.832187     0.9384375    1.727188          42.03125
13  9.904286     1.0200000    1.665714          46.69286
14  9.456818     0.8704545    1.646818          39.96909
15  8.542857     0.8635714    1.782857          45.12571
```

FIGURA 31 VALORI MEDI NEI CLUSTER DEGLI INDICI DI BILANCIO

Questi centri spiegano bene la segmentazione delle diverse fasce di imprese nel settore.
Si può notare come gli indici di bilancio medi di ciascun cluster (quelli che abbiamo definito come gli indici di performance) rimangano tutti all'interno di intervalli piuttosto ristretti.
Questo fatto non si nota nelle statistiche descrittive delle variabili perché siamo passati dai dati riferiti alla filiera (1253 osservazioni) ai dati del settore studiato (272 osservazioni).
Nel settore 271100 il ROI si aggira tra il 7% e il 14%, il turnover ratio tra il 90% e il 113%, l'indice di liquidità tra 1.3% e 1.8%, e l'indice di indipendenza finanziaria tra il 39% e il 45%.
Si può quindi notare quanto le differenze di performance, e quindi di strategia, tra imprese appartenenti a diversi cluster siano difficili da valorizzare.

Vediamo che il cluster 5 è quello di dimensioni maggiori, almeno per numero di dipendenti e fatturato, seguito dal cluster 15 e dal cluster 7.
Il cluster 9, l'unico con EBITDA negativo e con ROI e turnover ratio tra i più bassi, presenta probabilmente le performance peggiori.

Vediamo, ad esempio, che le imprese del cluster 6 presentano ROI, indice di liquidità e indice di indipendenza finanziaria superiori rispetto a quelle del cluster 10, eppure le due imprese hanno più o meno le stesse dimensioni, in quanto numero di dipendenti e fatturato non differiscono di molto.
Possiamo già presumere che il cluster 6 sia più perfomante tra i due, ma l'individuazione delle cause può derivare unicamente dalle nostre interpretazioni degli altri scostamenti tra i due valori medi di bilancio. In particolare, il cluster 6 versa più imposte del cluster 10, investe maggiormente in immobilizzazioni materiali, sostiene minori costi di ricerca e pubblicità e in diritti di brevetto.
Per contro, vediamo che il cluster 6 sostiene costi per godimento di beni di terzi molto superiori rispetto a quelli del cluster 10, nonché una spesa in servizi maggiore e una spesa in materie prime dimezzata. Si può pertanto presumere che il cluster 6, a dispetto del cluster 10, avendo le stesse dimensioni, predilige una politica di outsourcing, tale da rendere le sue attività produttive maggiormente specializzate e meno diversificate. Questa differenza pone le imprese del primo su un piano di superiorità strategica e di efficienza mediamente superiore a quello delle imprese del secondo.

5.8 Ranking dei cluster

Per capire quale cluster sia il migliore, dovremmo simultaneamente tenere conto del punteggio che ciascun cluster ha ottenuto su ciascuna delle variabili di valutazione: ROI, Turnover Ratio, Indice di Indipendenza Finanziaria e Indice di Liquidità.

Possiamo utilizzare il Metodo Borda per calcolare rapidamente un punteggio univoco da associare a ciascun cluster.

```
> BordaRank(centri[,indici]) ##punteggio ottenuto da ciascun cluster
 1  2  3  4  5  6  7  8  9 10 11 12 13 14 15
24 27 34 45 30 38 42 37 25 21 27 32 44 25 29
> round(16 - rank(BordaRank(centri[,indici]))) ##rango finale ottenuto da ciascun cluster
 1  2  3  4  5  6  7  8  9 10 11 12 13 14 15
14 10  6  1  8  4  3  5 12 15 10  7  2 12  9
```

FIGURA 32 PUNTEGGIO E RANGO DEL METODO BORDA APPLICATO SUGLI INDICI MEDI DI BILANCIO DEI CLUSTER

Vediamo che il cluster complessivamente più performante, almeno con rifermento agli indici considerati, risulta essere il numero 4, seguito dal 13 e dal 7. I cluster più performanti non sono quindi quelli di dimensioni maggiori.
Il cluster 9 non finisce in fondo alla classifica, come si sarebbe potuto presumere osservando il suo EBITDA e i suoi indici di performance, ma è comunque tra i peggiori.
Con riferimento all'esempio precedente, possiamo notare che il cluster 6 presenta indici di redditività e solidità finanziaria complessivamente superiori rispetto a quelli del cluster 10.
Ecco i centri dei cluster ordinati per il punteggio Borda ricevuto:

	Group.1	Ricavi	SalariStipendi	TotImposte	Costi_ricerca_pubb	Diritti_brevetto	Dipendenti	EBITDA
4	4	28072.53	3647.684	192.8421	77.57895	336.57895	91.94737	3321.3684
13	13	24933.57	3473.857	445.8571	0.00000	54.14286	102.42857	3302.2857
7	7	74781.32	4639.789	1340.7895	417.10526	229.89474	110.05263	6508.6842
6	6	19633.08	2459.125	447.5000	37.25000	19.65000	74.75000	2617.6500
8	8	33453.86	2272.714	162.8571	15.85714	10.85714	60.57143	2959.4286
3	3	28524.07	2630.000	448.2667	104.73333	29.26667	73.20000	3276.4667
12	12	20786.09	2382.000	417.9375	56.87500	145.96875	67.06250	2357.7500
5	5	131033.23	20324.154	1074.3077	43.92308	185.38462	505.61538	9371.8462
15	15	34199.57	4856.000	1293.7857	1270.64286	524.21429	113.14286	7785.5000
2	2	38673.90	5757.500	1229.1000	35.70000	517.70000	152.20000	6494.6000
11	11	26846.06	3606.312	236.1875	150.68750	13.68750	98.06250	1954.3750
9	9	29358.48	2135.522	-154.2174	51.69565	123.69565	59.21739	-982.7826
14	14	29649.05	3571.727	331.3182	65.09091	279.81818	85.77273	2093.3182
1	1	56118.12	2706.875	75.3750	1.50000	80.12500	77.12500	486.7500
10	10	20669.76	2516.879	335.6061	143.57576	47.30303	69.21212	2200.4545

	CostiMateriePrime	CostiServizi	CostiGodimentoBeniDiTerzi	CostiImpiantoAmpl	ImmoMateriali	rango
4	15678.37	5261.000	275.84211	0.0000000	7311.211	1
13	13755.86	4234.143	311.42857	0.0000000	5526.429	2
7	55831.11	6992.211	744.21053	7.7894737	10265.368	3
6	11538.02	3478.800	365.05000	47.2750000	4244.900	4
8	26023.43	2738.571	28.42857	0.0000000	7148.286	5
3	17368.13	5166.333	440.20000	2.8666667	3380.533	6
12	12338.03	3753.781	327.81250	0.8125000	4552.469	7
5	81212.69	15988.308	963.23077	0.6923077	38339.846	8
15	12277.71	5932.500	708.35714	55.7142857	6414.429	9
2	20397.70	5391.100	292.40000	39.6000000	9373.000	10
11	14568.56	6719.062	230.68750	0.5625000	6075.750	11
9	21178.70	4105.652	235.13043	0.2173913	5239.783	12
14	18434.68	5331.455	450.50000	3.0454545	4800.864	13
1	49978.62	5001.000	437.12500	0.1250000	1955.750	14
10	12663.21	3372.000	209.00000	10.0303030	3936.788	15

FIGURA 33 VALORI MEDI DELLE VARIABILI STRATEGICHE DEI CLUSTER ORDINATI IN BASE AL PUNTEGGIO BORDA ASSEGNATO

Vediamo che non c'è una singola variabile direttamente correlata con il punteggio Borda (nessuna variabile cresce o si riduce sempre o quasi al crescere del rango finale), ma che il risultato complessivo degli indici di bilancio dipende da una combinazione dei singoli volumi di investimento, oltre che da altri eventuali fattori ambientali o specifici non considerati.

5.9 Interpretazione dei risultati

L'analisi dei cluster ha mostrato quanto siano sottili le differenze tra le diverse fasce della popolazione, e quanto risulti complesso ottenere una sola strategia di esempio.
I cluster di benchmark sono, in ordine di performance, il 4, il 13 e il 7. Notiamo che le ridotte dimensioni in termini di fatturato e numero di dipendenti dei cluster 13, 6 e 8 ci forniscono una strategia di esempio nel caso in cui volessimo fondare un'impresa di piccole dimensioni, per provare a sondare il mercato senza assumerci un rischio finanziario ingente. Le dimensioni del cluster 7 riflettono invece quelle di imprese di grandi dimensioni che conseguono risultati ottimali.
I loro valori medi costituiscono punti di riferimento qualora volessimo costituire una nuova azienda operante nel settore. D'altro canto, gli stessi ci offrono dei benchmark di riferimento nel caso intendessimo valutare l'acquisizione di un'azienda già avviata. Si tratta quindi non di un risultato univoco, ma di diversi risultati che consentono di soddisfare diversi scopi.
Le dimensioni del cluster 4 si avvicinano alle dimensioni medie di tutti gli altri cluster, e i suoi valori medi possono indicare una strategia di benchmark complessiva.
Se vorremo avviare una nuova impresa sul mercato, converrà quindi assumere intorno ai 100 dipendenti, investire (in euro) circa 15 milioni in materie prime, 5 milioni in servizi, e 7 milioni in immobilizzazioni materiali, nonché lasciare ampi spazi agli investimenti in brevetti (non oltre 400k) e in outsourcing (tra un minimo di 250k e non oltre gli 800k).
Naturalmente i valori ottenuti possono essere solo approssimazioni, ma la loro significatività cresce al crescere della popolazione nei cluster.
Includendo diverse variabili decisionali e di performance, è possibile identificare combinazioni strategiche più idonee al raggiungimento degli obiettivi considerati.

6. Un secondo esempio: dove localizzare l'azienda?

Dal dataset originario, della filiera produttiva, possiamo ricavare altre informazioni utili al decision making di chi intende entrare nel mercato, fondando una nuova impresa o mediante acquisizioni.
Vogliamo infatti comprendere quale posizione, approssimativamente, in Italia, può favorire il rendimento dell'impresa che opera nella fabbricazione di motori o trasformatori elettrici.
In questo caso l'obiettivo consiste nell'individuare la latitudine e la longitudine media delle province che offrono condizioni esterne favorevoli all'ingresso dell'impresa nel mercato.
Queste condizioni possono essere estremamente specifiche e riferite unicamente al settore analizzato, oppure possono seguire principi generali. Ad esempio, è condivisa l'dea che le imprese tendono a posizionarsi in aree geografiche dove vi è un'alta concentrazione di scuole e competenze[87]. La vicinanza a università e istituti professionali favorisce lo sviluppo del business e aiuta le imprese a ottenere maggior potere contrattuale sul mercato del lavoro[88].
Sulla localizzazione geografica si fonda anche il potere di mercato dei distretti industriali, che attraverso competenze, collaborazioni con le istituzioni, e una supply chain con bassi costi di trasporto e transazione, riescono a mantenere una posizione dominante sul territorio.
I distretti industriali possono però rappresentare un'arma a doppio taglio: le imprese che vi si stabiliscono rischiano un completo annientamento da parte di eventuali new entry in grado di produrre a prezzi inferiori e di fidelizzare i medesimi partner[89].
Si potrebbero considerare molte altre dimensioni caratteristiche del territorio, come il numero di istituti scolastici in grado di fornire le competenze richieste nel settore, o il peso che possono costituire altri ostacoli, come barriere legali, opinione pubblica o livello di criminalità organizzata.
La condizione naturale per poter integrare l'analisi di queste dimensioni è rappresentata dalla concreta possibilità di accedere a relative misurazioni quantitative (es. numero di scuole, numero di arresti, condizioni climatiche ecc.) con un livello di dettaglio geografico pari a quello considerato (a livello di provincia, di regione, di città o di paese).
Vogliamo in questo caso separare le province italiane più simili e vicine fra loro, per segmentare il settore dal punto di vista geografico. Si potrebbe anche ricorrere ai sistemi locali del lavoro, e svolgere un'analisi fondata su conoscenze a priori dei principali distretti industriali.

[87] Camuffo, A., & Grandinetti, R. (2011). I distretti industriali come sistemi locali di innovazione. *Sinergie Italian Journal of Management*, (69), 33-60.
[88] Carloni, M., Ciarrocchi, A., & Micozzi, A. (2020). La vicinanza all'Università? Un'opportunità. Le scelte di localizzazione delle start-up innovative italiane. *L'industria*, *41*(2), 269-289.
[89] Ricciardi, A. (2013). I distretti industriali italiani: recenti tendenze evolutive (Italian industrial districts: recent evolutionary trends). *Sinergie Italian Journal of Management*, *31*(May-Aug), 21-58.

6.1 Il dataset delle province

Per adottare il metodo già descritto, costruiamo un nuovo dataset, contenente le principali province italiane sulle righe (86 osservazioni), 2 variabili di valutazione strategica e 3 variabili di performance. Dovendo scegliere dove localizzare l'azienda, le variabili definizione strategica, sulle quali saremo chiamati a prendere delle decisioni, non potranno che essere la latitudine e la longitudine.
In questo caso, per ogni provincia, assumiamo come latitudine e longitudine i valori medi delle latitudini e delle longitudini di tutte le città presenti nella provincia.
Le variabili di performance le possiamo ricavare dal dataset di filiera. In questo caso consideriamo, per ogni provincia:

- il numero di imprese potenzialmente concorrenti, ovvero già operanti nel settore in cui si vuole entrare
- il numero di imprese potenzialmente partner, ovvero operanti in altri settori a monte o a valle della medesima filiera produttiva
- il numero di lavoratori dipendenti impiegati nella filiera: se in una certa provincia vi è una solida presenza di lavoratori operanti nella filiera, il mercato del lavoro presenterà condizioni favorevoli e la ricerca di personale dipendente sarà meno costosa e potrà offrire più competenze per l'azienda entrante.

```
> summary(dati_province)
      lat              lon          num_partner     num_competitor    num_workers
 Min.   :36.89   Min.   : 7.338   Min.   :  0.00   Min.   : 0.000   Min.   :    2.0
 1st Qu.:41.37   1st Qu.:10.330   1st Qu.:  2.00   1st Qu.: 0.000   1st Qu.:   46.5
 Median :43.91   Median :12.137   Median :  5.00   Median : 1.000   Median :  278.5
 Mean   :43.08   Mean   :12.171   Mean   : 11.29   Mean   : 3.279   Mean   :  967.8
 3rd Qu.:45.18   3rd Qu.:13.743   3rd Qu.: 10.75   3rd Qu.: 3.000   3rd Qu.:  789.2
 Max.   :46.71   Max.   :18.178   Max.   :173.00   Max.   :49.000   Max.   :12402.0
```

FIGURA 34 STATISTICHE DESCRITTIVE DEL DATASET DELLE PROVINCE

Naturalmente è possibile utilizzare altri indici diversi da quelli qui proposti, purché coerenti con lo scopo dell'indagine. Possiamo visualizzare rapidamente la densità delle variabili di valutazione per ciascuna provincia italiana attraverso delle heatmap cartografiche:

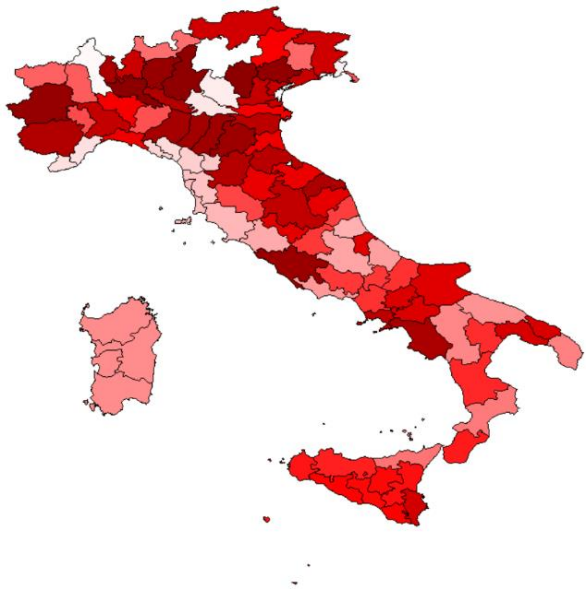


FIGURA 35 N° COMPETITOR DEL SETTORE ATECO 271100: FABBRICAZIONE DI GENERATORI, MOTORI E TRASFORMATORI ELETTRICI

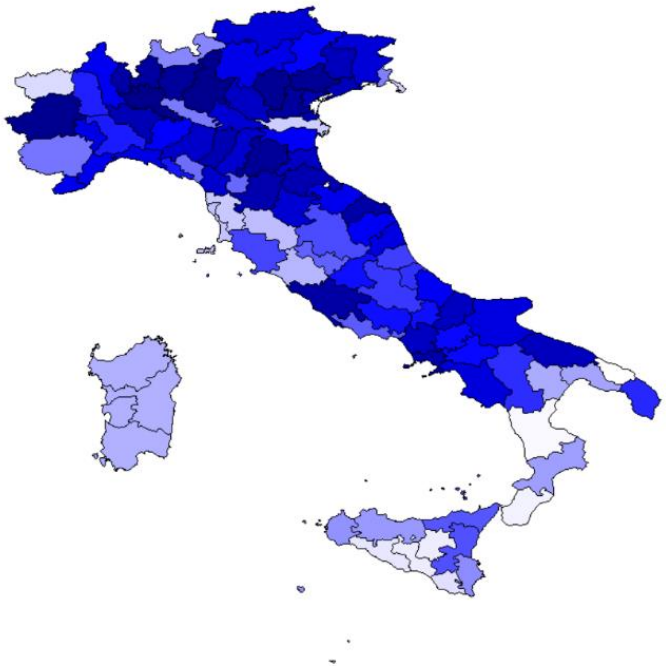


FIGURA 36 N° POTENZIALI IMPRESE PARTNER IMPEGNATE NELLA FILIERA DI FABBRICAZIONE DI APPARECCHIATURE ELETTRONICHE

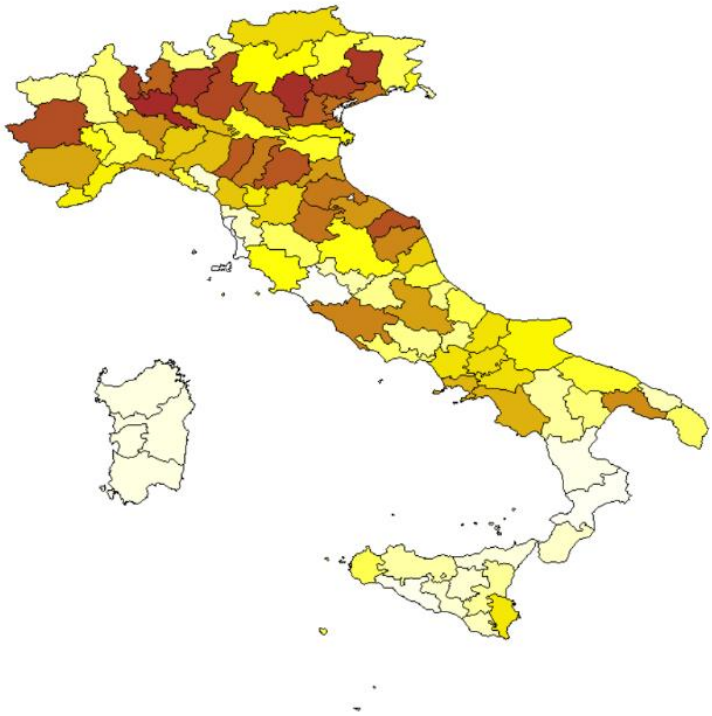
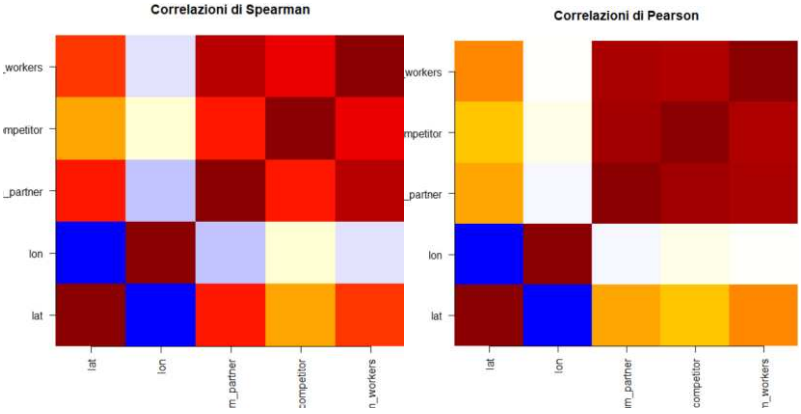
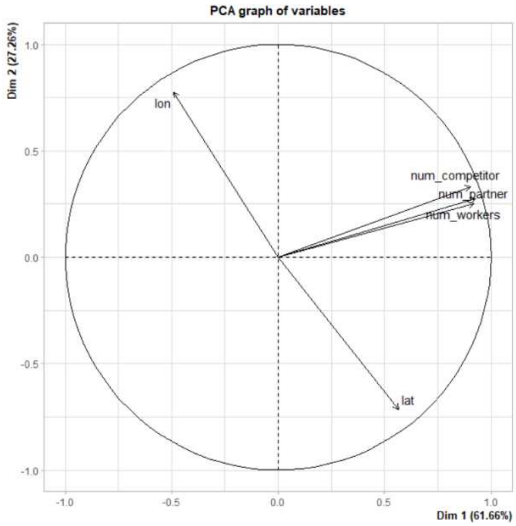


FIGURA 37 N° LAVORATORI DIPENDENTI COMPLESSIVAMENTE IMPEGNATI NELLA FILIERA DI FABBRICAZIONE DI APPARECCHIATURE ELETTRONICHE

6.2 Analisi bivariata e fattoriale






```
> summary(pca)
Importance of components:
                          PC1    PC2     PC3     PC4     PC5
Standard deviation     1.7559 1.1676 0.57538 0.35953 0.30566
Proportion of Variance 0.6166 0.2727 0.06621 0.02585 0.01869
Cumulative Proportion  0.6166 0.8892 0.95546 0.98131 1.00000
```

FIGURA 38 RISULTATI DELL'ANALISI BIVARIATA E FATTORIALE DELLE PROVINCE

Il biplot e i coefficienti di correlazione forniscono quattro informazioni.

In Italia, come si può intuire, la latitudine e la longitudine sono negativamente correlate, anche se debolmente. Il numero di dipendenti, il numero di imprese operanti nel settore 271100 e il numero di imprese con codice Ateco 27**** sono positivamente correlati, in quanto riflettono insieme l'intensità dell'attività della filiera nello stesso territorio.

Si può anche notare che la longitudine non risulta significativamente correlata con il numero di imprese e di dipendenti della filiera e del settore, a differenza della latitudine, che risulta molto correlata con la filiera, e meno correlata con il settore, anche se in modo significativo. A fronte di queste differenze, si potrebbe considerare di raccogliere due o tre componenti principali. Il metodo del gomito ci fornisce la soluzione, con una variabilità totale spiegata del 95.55%.

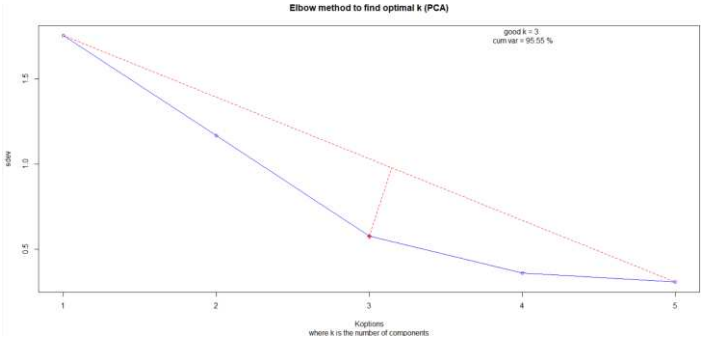


FIGURA 39 ELBOW METHOD PER LA PCA DELLE PROVINCE

6.3 Distanze nei dati

Possiamo utilizzare la matrice delle distanze per visualizzare rapidamente i dati:

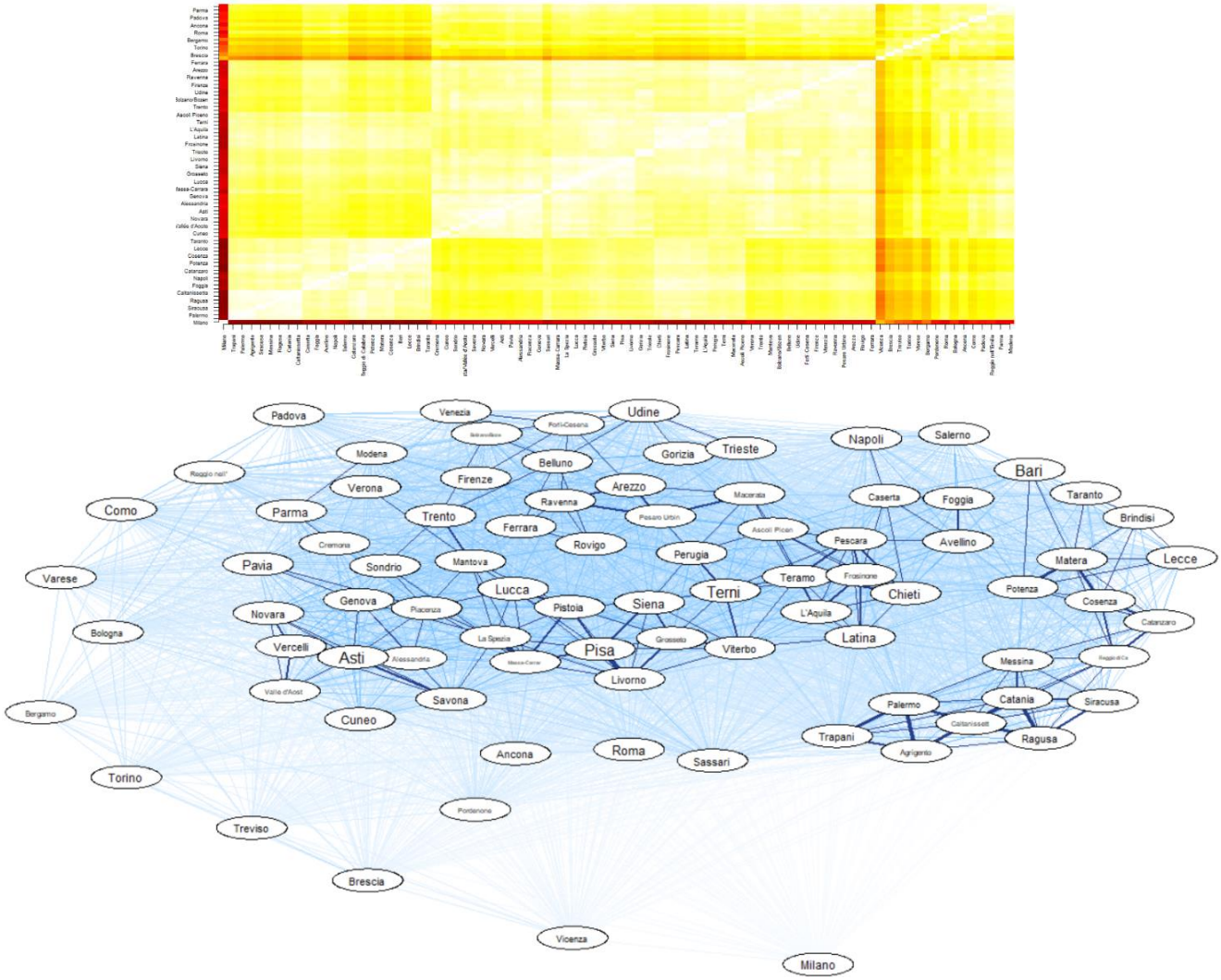


FIGURA 40 MATRICE DELLE DISTANZE TRA LE PROVINCE ITALIANE CON RIFERIMENTO AL SETTORE ATECO 271100

Le osservazioni non sembrano aggregarsi una volta sola, per cui possiamo presumere, almeno a livello campionario, un numero di cluster maggiore di 1.
Vediamo che le province di Milano e Vicenza rappresentano probabilmente degli outlier. In questo contesto si decide di non rimuoverle, per poter utilizzare una rappresentazione cartografica dei cluster. Possiamo aspettarci che alle due province venga attribuito un gruppo a sé stante.

6.4 Cluster di province

In questo caso si ha un'equivalenza tra le silhouette medie per gli algoritmi di clustering più comuni (gerarchico, kmeans e medoids), per livelli di k maggiori di 2:

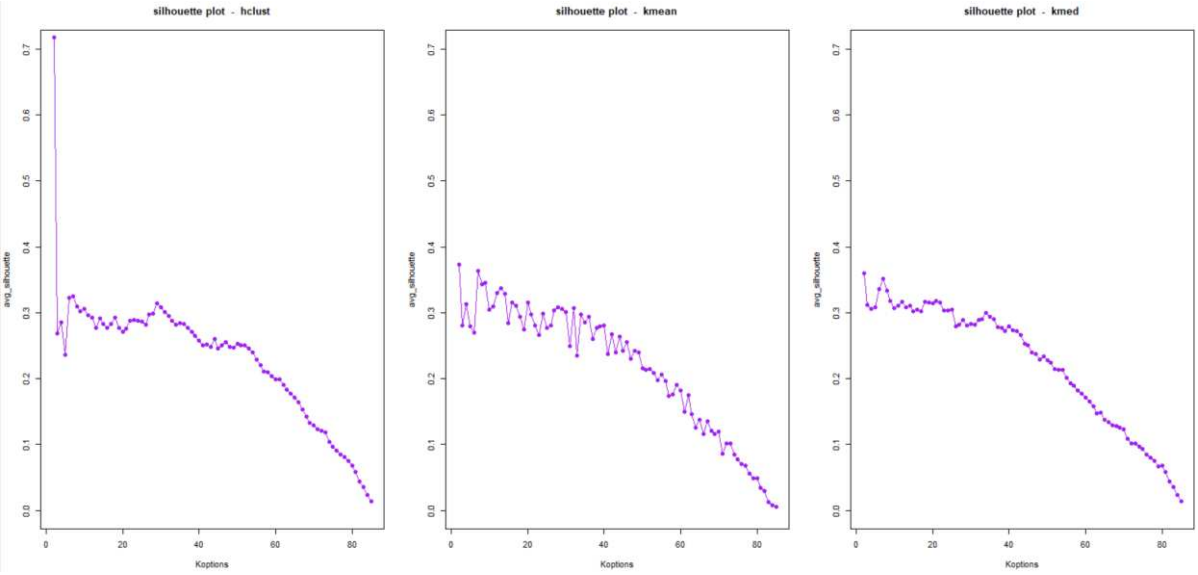


FIGURA 41 SILHOUETTE DEL CLUSTERING GERARCHICO E PARTIZIONALE SULLE PROVINCE PER OGNI NUMERO DI GRUPPI

Per individuare un numero ottimale di k, non avendo una conoscenza a priori del numero di gruppi, ci possiamo affidare agli indici di dispersione interna ricavati dal kmeans clustering, ancora con il metodo del gomito:

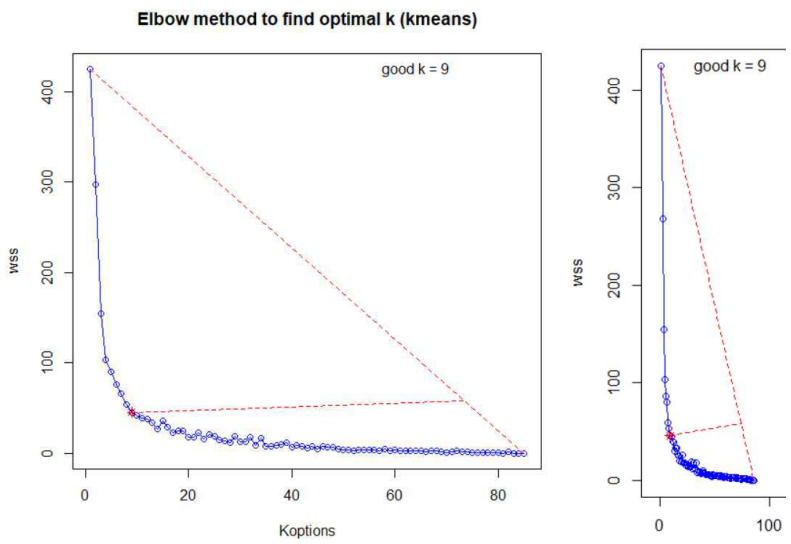


FIGURA 42 METODO ELBOW PER L'INDIVIDUAZIONE DEL NUMERO DI CLUSTER GEOGRAFICI CON KMEANS (NOTA: NEL GRAFICO DI SINISTRA LE SCALE DEGLI ASSI DIFFERISCONO)

In questo caso, avendo un dataset di sole 86 righe, sembra ragionevole ripartire la popolazione in 9 cluster. Notiamo che due osservazioni rimangono outlier, che non sono rimossi al fine di mantenere una visualizzazione cartografica dei cluster.

$$k \approx \sqrt{\frac{86}{2}} = 6.55$$

Anche il gomito calcolato sulle altezze dei rami del dendrogramma del clustering gerarchico (con distanza euclidea e complete linkage) conferma la scelta:

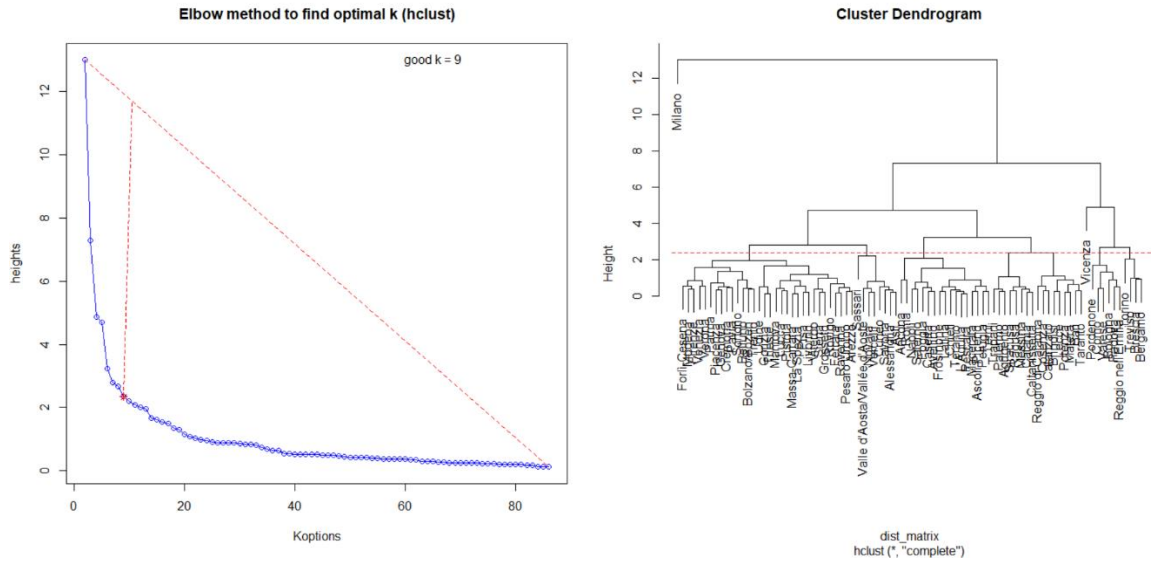


FIGURA 43 METODO ELBOW PER IL NUMERO DI CLUSTER GEOGRAFICI CON IL DENDROGRAMMA (EUCLIDEAN DISTANCE, COMPLETE LINKAGE) (NOTA: LE SCALE DEGLI ASSI DIFFERISCONO)

Va evidenziato che la provincia di Sassari (la Sardegna in generale) si presta anch'essa a considerarsi un outlier, come si può notare dal dendrogramma, in ragione della localizzazione geografica.
Ecco una rappresentazione grafica dei cluster costruiti dal kmedoids, adottando tre componenti principali:

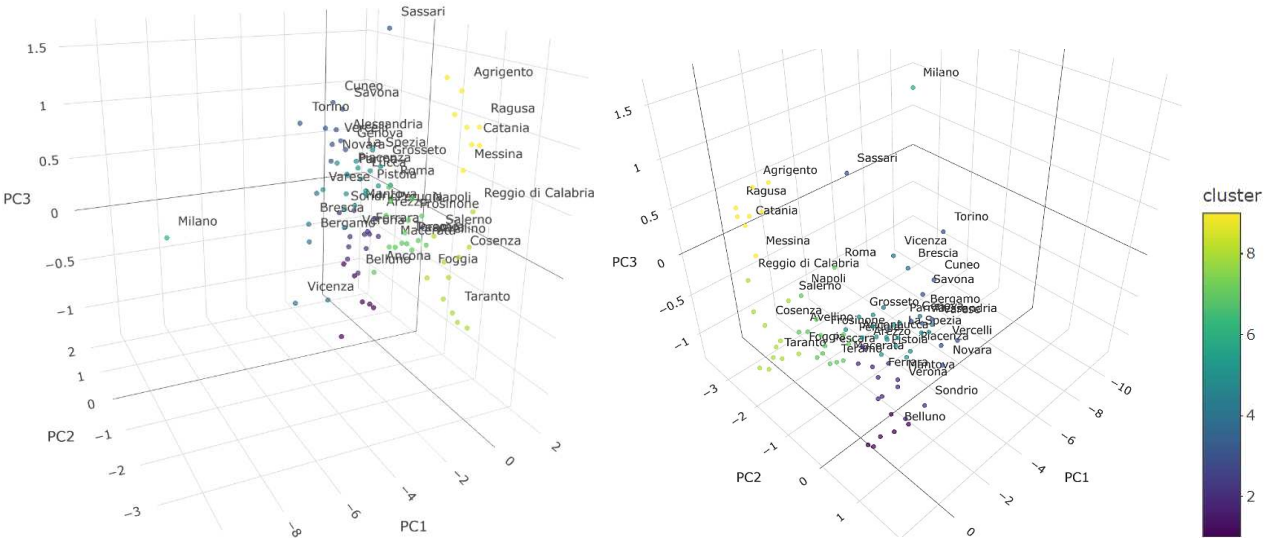


FIGURA 44 CLUSTER DELLE PROVINCE OTTENUTI CON KMEDOIDS (43 ETICHETTE SU 86 PUNTI)

Si riportano di seguito i valori medi di ciascun cluster:

```
> centri
  Group.1      lat       lon num_partner num_competitor num_workers
1       1 46.13844 12.769197    6.166667       1.666667   1127.0000
2       2 44.93458 11.644088   10.250000       1.916667    784.5833
3       3 44.50555  8.161007    9.100000       2.200000    557.1000
4       4 45.53215 10.652655   42.666667      14.833333   4775.1667
5       5 44.10954 10.415650    7.647059       2.176471    553.8235
6       6 45.36911  9.316859  173.000000      49.000000  12402.0000
7       7 42.19841 13.564152    8.000000       1.933333    555.7333
8       8 40.19230 16.537099    3.090909       1.363636    193.0909
9       9 37.50924 14.135614    0.875000       1.000000     69.7500
```

FIGURA 45 CENTRI DEI CLUSTER DELLE PROVINCE

Ed ecco la numerosità di ciascun cluster: notiamo la presenza dei due noti outlier nei cluster 4 e 5.

```
> table(kmedoid$clustering) ##numerosità di ciascun cluster

 1  2  3  4  5  6  7  8  9
 6 12 10  6 17  1 15 11  8
```

FIGURA 46 NUMEROSITÀ DEI CLUSTER DELLE PROVINCE

Milano (cluster 6) costituisce un cluster a sé stante, trattandosi di una provincia nella quale sono presenti distretti industriali sovradimensionati rispetto alla media nazionale, mentre la numerosità degli altri cluster appare equamente distribuita.
La silhouette media risulta del 32%.
Il cluster di Milano presenta silhouette nulla, in quanto ha numerosità pari a 1.

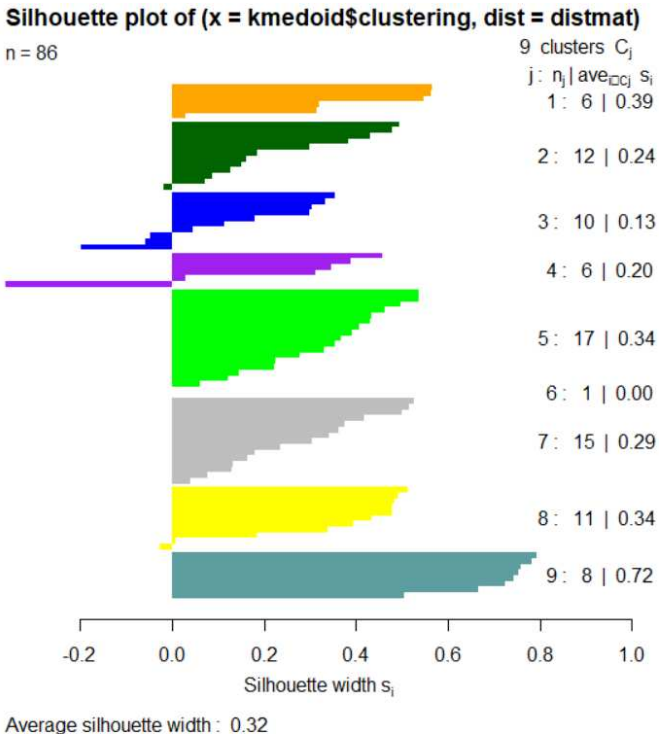


FIGURA 47 SILHOUETTE DEI CLUSTER DELLE PROVINCE

6.5 Localizzazione ottimale

La classifica del Metodo Borda premierà maggiormente i cluster, e quindi le province, con il maggior numero di dipendenti nella filiera e di potenziali aziende partner, e con il minor numero di potenziali aziende competitor.
Essendo la variabile num_competitor un valore che si vorrebbe minimizzare, la si può sottrarre al suo massimo per ottenere un ordinamento crescente dei valori preferibili.
Ogni punteggio, trasformato come (n-r), viene poi sommato a tutti gli altri.
Prima di assegnare un punteggio Borda a ciascun cluster, possiamo assegnarlo a ciascuna provincia, in modo tale da poter visualizzare su una cartina heatmap la posizione nazionale di ciascuna.

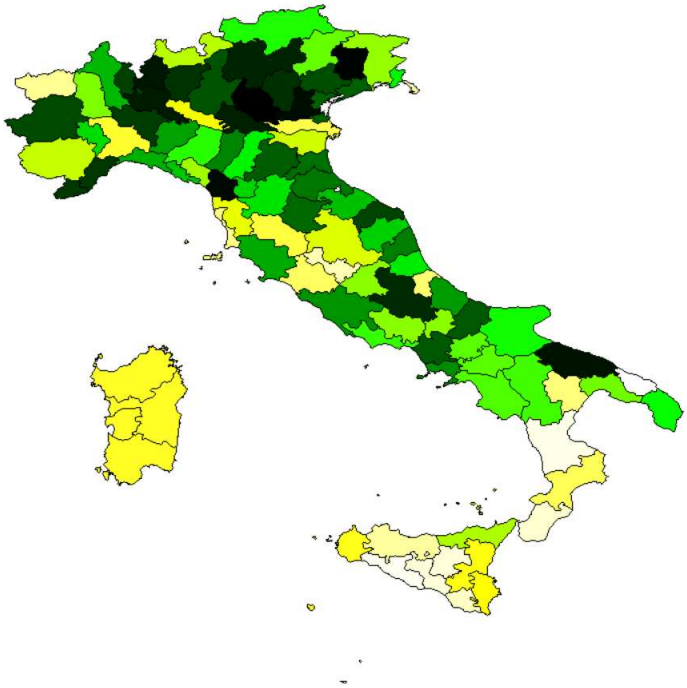
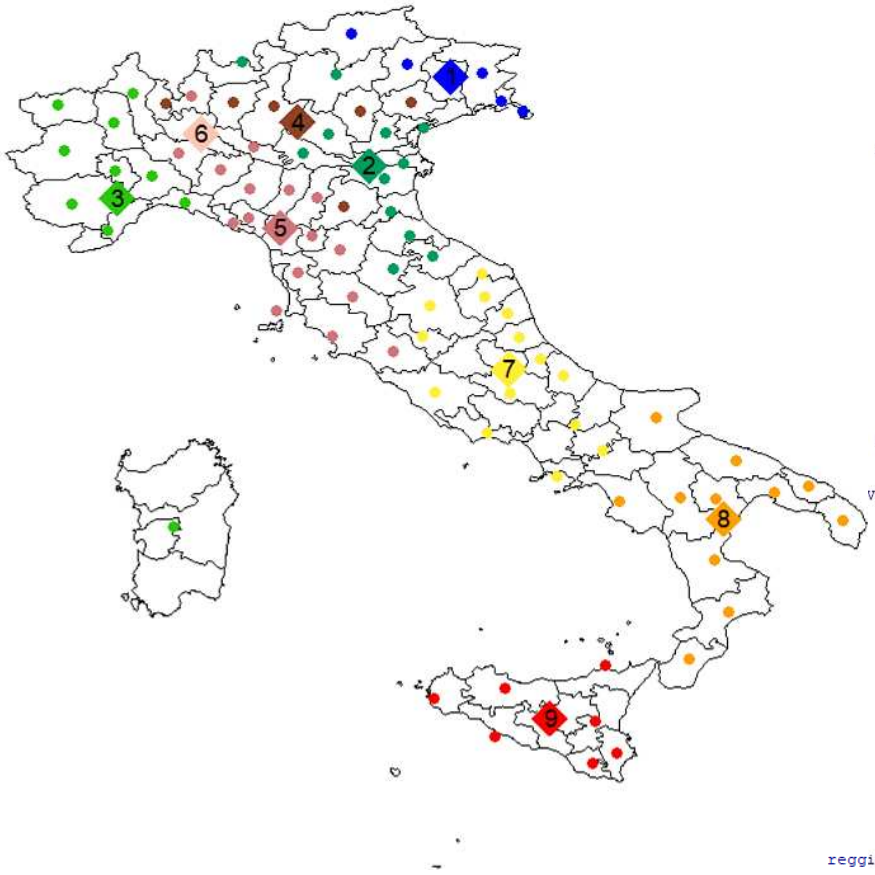


FIGURA 48 PUNTEGGIO BORDA OTTENUTO DALLE SINGOLE PROVINCE ITALIANE

Vediamo che la provincia di benchmark non è Milano, come si sarebbe potuto supporre in base alle dimensioni del suo mercato, ma vi sono diverse province con un punteggio vicini o superiori al suo. In particolare, emergono Torino, Imperia, L'Aquila, Lucca, Ancona e Bari, oltre all'area evidenziata da colori più scuri compresa fra Milano e Pordenone.
La localizzazione ottima non necessariamente coincide con la provincia con il ranking maggiore, in quanto vi possono essere filiere produttive localizzate su due o più province adiacenti, ma potrebbe coincidere con la latitudine e la longitudine media dei cluster migliori. Abbiamo incluso le variabili latitudine e longitudine proprio al fine di mantenere una concentrazione territoriale dei cluster.
Semplicemente osservando la heatmap, si può intuire che tale area di ottimo si posizioni tra Milano e Vicenza. Di seguito la latitudine e la longitudine media di ciascun cluster.



provincia	cluster	provincia	cluster
Bolzano/Bozen	1	Modena	5
Belluno	1	Massa-Carrara	5
Udine	1	La Spezia	5
Pordenone	1	Lucca	5
Gorizia	1	Firenze	5
Trieste	1	Pistoia	5
Sondrio	2	Pisa	5
Trento	2	Livorno	5
Venezia	2	Siena	5
Verona	2	Grosseto	5
Padova	2	Viterbo	5
Mantova	2	Milano	6
Rovigo	2	Ancona	7
Ferrara	2	Perugia	7
Ravenna	2	Macerata	7
Forlì-Cesena	2	Ascoli Piceno	7
Pesaro Urbino	2	Terni	7
Arezzo	2	Teramo	7
Novara	3	Frosinone	7
Vallée d'Aoste	3	L'Aquila	7
Vercelli	3	Pescara	7
Torino	3	Chieti	7
Alessandria	3	Roma	7
Asti	3	Caserta	7
Cuneo	3	Latina	7
Genova	3	Avellino	7
Savona	3	Napoli	7
Sassari	3	Foggia	8
Brescia	4	Bari	8
Varese	4	Potenza	8
Treviso	4	Brindisi	8
Bergamo	4	Salerno	8
Vicenza	4	Matera	8
Bologna	4	Taranto	8
Como	5	Lecce	8
Cremona	5	Cosenza	8
Pavia	5	Catanzaro	8
Piacenza	5	Reggio di Calabria	8
Parma	5	Messina	9
reggio nell'Emilia	5	Palermo	9

FIGURA 49 LATITUDINI E LONGITUDINI MEDIE DEI CLUSTER DELLE PROVINCE ITALIANE: I ROMBI INDICANO LA LATITUDINE E LA LONGITUDINE MEDIA DI CIASCUN CLUSTER

Applichiamo quindi il Metodo Borda ai cluster appena osservati

```
> BordaRank(centri2[,indici]) ##punteggio cluster
 1  2  3  4  5  6  7  8  9
17 19 14 18 11 19 14 12 11
```

FIGURA 50 PUNTEGGIO BORDA OTTENUTO DAI CLUSTER DELLE PROVINCE ITALIANE

Dai risultati del ranking, e tenendo conto dell'anomalia di un outlier, possiamo assumere che il cluster più performanti risultino essere il cluster 2 e il cluster 4, in terza posizione nella classifica, dopo Milano (cluster 6). Dell'area più scura, posizionata nel cuore industriale del Nord-Italia, fanno parte province assegnate in larga misura al cluster 4, in secondo luogo al cluster 2.

Vediamo che il Nord Italia, nelle province intorno a Vicenza, presenta le condizioni più favorevoli per la costituzione e lo sviluppo di un'impresa di questo tipo. Inoltre, il cluster 1, intorno a Pordenone, presenta ottime performance, distanziandosi con un solo punto in meno rispetto al cluster 4.

L'utilizzo di altre variabili di valutazione consentirebbe di definire meglio la natura delle condizioni ambientali che si vorrebbero raggiungere, e di conseguenza di ottenere risultati più vicini alle esigenze dell'impresa.

Naturalmente, avendo adottato la latitudine e la longitudine media di ciascuna provincia italiana come riferimento per le singole osservazioni, le coordinate dei centri dei cluster sono da ritenersi approssimazioni dell'area geografica in cui conviene posizionare l'azienda.

7. Conclusioni

Il metodo di analisi presentato si presenta molto utile e aperto.
È da considerarsi aperto sia rispetto alla tipologia di risultati che si vuole raggiungere sia dal punto di vista dei dati cui può essere applicato. I problemi posti nelle domande di ricerca iniziali hanno incontrato soluzioni coerenti con le aspettative.
L'analisi delle strategie e dei benchmark delle imprese operanti nel settore ha fornito informazioni sul volume di investimenti distribuito sulle diverse risorse aziendali, da parte di quelle imprese che mantengono indici di performance elevati.
L'analisi della distribuzione geografica della filiera produttiva ha consentito invece di individuare le aree geografiche italiane nelle quali un'impresa entrante può incontrare condizioni favorevoli per la propria crescita sul mercato.
Tali informazioni, ricavate da entrambe le analisi, si prestano a una varietà di utilizzi, da parte di numerose categorie di imprese e operatori economici.
Possono essere informazioni utili per valorizzare la presenza di distretti industriali sul territorio, o per individuare potenziali punti di partenza per fondare dei nuovi distretti, date le condizioni geo economiche tipicamente favorevoli per la specifica filiera.
Se si intendesse entrare in un nuovo settore, costituendo una nuova impresa o mediante l'acquisizione di una già operante in quel settore, le decisioni da prendere in merito alla strategia da adottare e alla localizzazione da scegliere riceverebbero un importante ausilio dagli algoritmi di clustering applicati alle due tipologie di dataset presentate, nonché ad altri dataset generati secondo i medesimi criteri.
Il principale limite dell'analisi è il quasi totale affidamento ai codici Ateco 2007. Sarebbe opportuno approfondire le analisi tramite classificazioni merceologiche differenti, come le classificazioni NACE o NAICS, e confrontare i risultati. Si potrebbero anche sperimentare diversi algoritmi di clustering sullo stesso dataset, e verificare la convergenza dei cluster. Infine, qualora si fosse in possesso di conoscenze di dominio specifiche del settore considerato, sarebbe possibile utilizzarle per decidere quali variabili includere e che tipo di analisi svolgere. In questa tesi si sono utilizzate variabili generiche ritenute importanti per qualsiasi categoria di impresa.

Bibliografia

[1] Benabdellah, A. C., Benghabrit, A., & Bouhaddou, I. (2019). A survey of clustering algorithms for an industrial context. *Procedia computer science*, *148*, 291-302.

Song, L., Dong, Y., Guo, Q., Meng, Y., & Zhao, G. (2023). An adaptive differential evolution algorithm with DBSCAN for the integrated slab allocation problem in steel industry. *Applied Soft Computing*, *146*, 110665.

Liu, G., Yang, J., Hao, Y., & Zhang, Y. (2018). Big data-informed energy efficiency assessment of China industry sectors based on K-means clustering. *Journal of cleaner production*, *183*, 304-314.

[2] Oyewole, G. J., & Thopil, G. A. (2023). Data clustering: Application and trends. *Artificial Intelligence Review*, *56*(7), 6439-6475.

[3] Mecca, G., Raunich, S., & Pappalardo, A. (2007). A new algorithm for clustering search results. *Data & Knowledge Engineering*, *62*(3), 504-522.

[4] Kuo, R. J., Ho, L. M., & Hu, C. M. (2002). Integration of self-organizing feature map and K-means algorithm for market segmentation. *Computers & Operations Research*, *29*(11), 1475-1493.

[5] Deng, Q., & Mei, G. (2009, August). Combining self-organizing map and k-means clustering for detecting fraudulent financial statements. In *2009 IEEE international conference on granular computing* (pp. 126-131). IEEE.

Sabau, A. S. (2012). Survey of clustering based financial fraud detection research. *Informatica Economica*, *16*(1), 110.

[6] Chong, D., & Zhu, H. H. (2012, December). Firm clustering based on financial statements. In *22nd Workshop on Information Technology and Information Systems (WITS'12)*.

Camuffo, A., & Grandinetti, R. (2011). I distretti industriali come sistemi locali di innovazione. *Sinergie Italian Journal of Management*, (69), 33-60.

[7] Enciclopedia Treccani - Dizionario di Economia e Finanza (2012) voce “settore”

Boggia, A., Carucci, A. M. M., & Filippello, R. istat working papers.

[8] Enciclopedia Treccani - Dizionario di Economia e Finanza (2012) voce “settore - settore industriale”

[9] Dalziel, M. (2007). A systems-based approach to industry classification. *Research Policy*, *36*(10), 1559-1574.

Schnabl, E., & Zenker, A. (2013). *Statistical classification of knowledge-intensive business services (KIBS) with NACE Rev. 2* (Vol. 25). Karlsruhe: Fraunhofer ISI.

[10] di ricerca Ceris-Cnr, G., Calabrese, G., Corio, G., Finardi, U., Manello, A., Ragazzi, E., ... & Saracco, P. Le caratteristiche socio-economiche dei cluster di imprese in Piemonte.

[11] Enciclopedia Treccani - Dizionario di Economia e Finanza (2012) voce “filiera”

[12] Zamora, E. A. (2016). Value chain analysis: A brief review. *Asian Journal of Innovation and Policy*, *5*(2), 116-128.

[13] Musso, F. (2012). *Innovazione nei canali di marketing*. Clueb.

[14] Grundy, T. (2006). Rethinking and reinventing Michael Porter's five forces model. *Strategic change*, *15*(5), 213-229.

[15] Skjott-Larsen, T. (2007). *Managing the global supply chain*. Copenhagen Business School Press DK.

[16] Sforzi, F. (Ed.). (1997). *I sistemi locali del lavoro 1991*. Roma: Istat.

Coppola, G., & Mazzotta, F. (2005). I sistemi locali del Lavoro in Italia: Aspetti teorici ed empirici.

Calafati, A. G., & Compagnucci, F. (2005). Oltre i sistemi locali del lavoro. *Economia Marche*, *24*(1), 51-76.

[17] Enciclopedia Treccani - Dizionario di Economia e Finanza (2012) voce “distretto industriale”

[18] Russo, M., Alboni, F., Sanginés, J. C., De Domenico, M., Mangioni, G., Righi, S., & Simonazzi, A. (2022). The Changing Shape of the World Automobile Industry: A Multilayer Network Analysis of International Trade in Components and Parts. *Institute for New Economic Thinking Working Paper Series*, (173).

[19] Balcet, G., & Enrietti, A. (1997). Regionalisation and globalisation in Europe: The case of Fiat Auto Poland and its suppliers. *Les Actes de GERPISA*, (20).

[20]Malik, S., Dedeoglu, V., Kanhere, S. S., & Jurdak, R. (2019, July). Trustchain: Trust management in blockchain and iot supported supply chains. In *2019 IEEE International Conference on Blockchain (Blockchain)* (pp. 184-193). IEEE.

Pal, K. (2023). Internet of Things Impact on Supply Chain Management. *Procedia Computer Science*, *220*, 478-485.

[21]Navaroni, M. (2020). La Position Analysis: la più Evoluta tra le Analisi di Bilancio. *Economia Aziendale Online-*, *11*(2), 133-144.

[22]Aithal, P. S. (2017). Industry Analysis–The First Step in Business Management Scholarly Research. *International Journal of Case Studies in Business, IT and Education (IJCSBE)*, *1*(1), 1-13.

[23]Aithal, P. S. (2017). ABCD Analysis as Research Methodology in Company Case Studies. *International Journal of Management, Technology, and Social Sciences (IJMTS)*, *2*(2), 40-54.

Saieva, V. (2012). International production relocation. *Economic Focus,(2)*, 1-26.

Nicolardi, V., & Marini, C. (2016). L’aggiornamento di strutture di Contabilità Nazionale disaggregate. In *Metodi e Analisi statistiche 2016* (pp. 131-146). Università degli Studi di Bari" Aldo Moro".

[24]Treccani - Enciclopedia delle scienze sociali (1992) voce “contabilità nazionale” - introduzione

[25]TANAKA, F. J. (2011). Applications of Leontief's input-output analysis in our economy.
[26]Round, J. (2003). Social accounting matrices and SAM-based multiplier analysis. *The impact of economic policies on poverty and income distribution: Evaluation techniques and tools*, *14*, 261-276.
[27]Giovanelli, L. (Ed.). (2006). *Contabilità dello stato e sistema europeo dei conti (SEC95) nella prospettiva comunitaria*. Giuffrè Editore.
[28]Santos, S. (2011). Constructing SAMs from the SNA.
[29]Mastrantonio (2018) https://www.istat.it/it/archivio/209141
[30]Parente, R. (2008). Co-evoluzione e cluster tecnologici. Roma: Aracne.
Smith, R. V. (2003). Industry cluster analysis: Inspiring a common strategy for community development. *Central Pennsylvania Workforce Development Corporation Report*, *296*.
Kim, H., Hwang, S. J., & Yoon, W. (2023). Industry cluster, organizational diversity, and innovation. *International Journal of Innovation Studies*, *7*(3), 187-195.
[31]Skiena, S. S. (2017). *The data science design manual* Cap. 10.5: clustering. Springer.
[32]Abbas, A., Prayitno, P., Nurkim, N., Prumanto, D., Dewadi, F. M., Hidayati, N., & Windarto, A. P. (2021, February). Implementation of clustering unsupervised learning using K-Means mapping techniques. In *IOP Conference Series: Materials Science and Engineering* (Vol. 1088, No. 1, p. 012004). IOP Publishing.
[33]Curea, S. C., Belascu, L., & Barsan, A. M. (2020). An Exploratory Study of Financial Performance in CEE Countries. *KnE Social Sciences*, 286-300.
[34]Russo, M., Pirani, E., & Paterlini, S. (2006). L'industria meccanica in Italia: una analisi cluster delle differenze territoriali. *Materiali di discussione*.
[35] Fanelli, R. M., & Felice, F. (2014). Un'applicazione dell'analisi multivariata e della convergenza non parametrica all'industria birraria italiana. *Italian Review of Agricultural Economics*, *69*(1), 7-30.
[36]Fung, G. (2001). A comprehensive overview of basic clustering algorithms.
[37]Fahad, A., Alshatri, N., Tari, Z., Alamri, A., Khalil, I., Zomaya, A. Y., ... & Bouras, A. (2014). A survey of clustering algorithms for big data: Taxonomy and empirical analysis. *IEEE transactions on emerging topics in computing*, *2*(3), 267-279.
[38]Sonagara, D., & Badheka, S. (2014). Comparison of basic clustering algorithms. *Int. J. Comput. Sci. Mob. Comput*, *3*(10), 58-61.
[39]Li, Y., & Wu, H. (2012). A clustering method based on K-means algorithm. *Physics Procedia*, *25*, 1104-1109.
[40]Batra, P. (2018). Comparative study of density based clustering algorithms.
Jahirabadkar, S., & Kulkarni, P. (2014). Algorithm to determine ε-distance parameter in density based clustering. *Expert systems with applications*, *41*(6), 2939-2946.
[41]Hahsler, M., Piekenbrock, M., & Doran, D. (2019). dbscan: Fast density-based clustering with R. *Journal of Statistical Software*, *91*, 1-30.
[42]Cheng, W., Wang, W., & Batista, S. (2018). Grid-based clustering. In *Data clustering* (pp. 128-148). Chapman and Hall/CRC.
Hireche, C., Drias, H., & Moulai, H. (2020). Grid based clustering for satisfiability solving. *Applied Soft Computing*, *88*, 106069.
[43]Gelbard, R., Goldman, O., & Spiegler, I. (2007). Investigating diversity of clustering methods: An empirical comparison. *Data & Knowledge Engineering*, *63*(1), 155-166.
[44]Rodriguez, M. Z., Comin, C. H., Casanova, D., Bruno, O. M., Amancio, D. R., Costa, L. D. F., & Rodrigues, F. A. (2019). Clustering algorithms: A comparative approach. *PloS one*, *14*(1), e0210236.
https://generativeai.pub/how-to-choose-the-right-clustering-algorithm-for-your-data-8f3ee24b9c16
[45]Skiena, S. S. (2017). *The data science design manual* Cap. 10.1: measuring distances. Springer.
[46]Rodrigues, É. O. (2018). Combining Minkowski and Chebyshev: New distance proposal and survey of distance metrics using k-nearest neighbours classifier. *Pattern Recognition Letters*, *110*, 66-71.
Singh, A., Yadav, A., & Rana, A. (2013). K-means with Three different Distance Metrics. *International Journal of Computer Applications*, *67*(10).
[47]Tsafrir, D., Tsafrir, I., Ein-Dor, L., Zuk, O., Notterman, D. A., & Domany, E. (2005). Sorting points into neighborhoods (SPIN): data analysis and visualization by ordering distance matrices. *Bioinformatics*, *21*(10), 2301-2308.
[48]Shetty, P., & Singh, S. (2021). Hierarchical clustering: a survey. *International Journal of Applied Research*, *7*(4), 178-181.
[49] Sokal, R. R., & Sneath, P. H. (1963). Principles of numerical taxonomy. *Principles of numerical taxonomy*.
Ehrlich, P. R. (1958). Problems of higher classification. *Systematic Zoology*, *7*(4), 180-184.
Yim, O., & Ramdeen, K. T. (2015). Hierarchical cluster analysis: comparison of three linkage measures and application to psychological data. *The quantitative methods for psychology*, *11*(1), 8-21.
[50]Mahmoud, M. (2012). *Genotype imputation based on discriminant and cluster analysis* (Master's thesis, Norwegian University of Life Sciences, Ås).
[51]https://r-graph-gallery.com/29-basic-dendrogram.html
[52]Forina, M., Armanino, C., & Raggio, V. (2002). Clustering with dendrograms on interpretation variables. *Analytica Chimica Acta*, *454*(1), 13-19.

[53]Boudaillier, E., & Hebrail, G. (1998). Interactive interpretation of hierarchical clustering. *Intelligent Data Analysis*, *2*(1-4), 229-244.
[54]Arbin, N., Suhaimi, N. S., Mokhtar, N. Z., & Othman, Z. (2015, December). Comparative analysis between k-means and k-medoids for statistical clustering. In *2015 3rd International Conference on Artificial Intelligence, Modelling and Simulation (AIMS)* (pp. 117-121). IEEE.
[55]Münz, G., Li, S., & Carle, G. (2007, September). Traffic anomaly detection using k-means clustering. In *Gi/itg workshop mmbnet* (Vol. 7, No. 9).
[56]Piech on kmeans (2013) https://stanford.edu/~cpiech/cs221/handouts/kmeans.html
[57]Kaufman, L., & Rousseeuw, P. J. (2009). *Finding groups in data: an introduction to cluster analysis*. John Wiley & Sons.
[58]Reynolds, A. P., Richards, G., de la Iglesia, B., & Rayward-Smith, V. J. (2006). Clustering rules: a comparison of partitioning and hierarchical clustering algorithms. *Journal of Mathematical Modelling and Algorithms*, *5*, 475-504.
Thinsungnoena, T., Kaoungkub, N., Durongdumronchaib, P., Kerdprasopb, K., & Kerdprasopb, N. (2015). The clustering validity with silhouette and sum of squared errors. *learning*, *3*(7).
[59]Ishchenko, I., Globa, L. S., Buhaienko, Y., & Liashenko, A. (2019). Approach to determining the number of clusters in a data set.
Zhang, Y., Mańdziuk, J., Quek, C. H., & Goh, B. W. (2017). Curvature-based method for determining the number of clusters. *Information Sciences*, *415*, 414-428.
Yilmaz, S., Chambers, J., Cozza, S., & Patel, M. K. (2019, November). Exploratory study on clustering methods to identify electricity use patterns in building sector. In *Journal of Physics: Conference Series* (Vol. 1343, No. 1, p. 012044). IOP Publishing.
[60]Kodinariya, T. M., & Makwana, P. R. (2013). Review on determining number of Cluster in K-Means Clustering. *International Journal*, *1*(6), 90-95.
Shi, C., Wei, B., Wei, S., Wang, W., Liu, H., & Liu, J. (2021). A quantitative discriminant method of elbow point for the optimal number of clusters in clustering algorithm. *Eurasip Journal on Wireless Communications and Networking*, *2021*(1), 1-16.
[61]Shalaby, M., Belal, N. A., & Omar, Y. (2021). Data clustering improves Siamese neural networks classification of Parkinson's disease. *Complexity*, *2021*, 1-9.
[62]Madhulatha, T. S. (2012). An overview on clustering methods.
[63]Çelik, M., Dadaşer-Çelik, F., & Dokuz, A. Ş. (2011, June). Anomaly detection in temperature data using DBSCAN algorithm. In *2011 international symposium on innovations in intelligent systems and applications* (pp. 91-95). IEEE.
Çelik, M., Dadaşer-Çelik, F., & Dokuz, A. Ş. (2011, June). Anomaly detection in temperature data using DBSCAN algorithm. In *2011 international symposium on innovations in intelligent systems and applications* (pp. 91-95). IEEE.
[64]Toller, M. Anomalies in Data.
Ester, M., Kriegel, H. P., Sander, J., & Xu, X. (1996, August). A density-based algorithm for discovering clusters in large spatial databases with noise. In *kdd* (Vol. 96, No. 34, pp. 226-231).
[65]Mohamad, I. B., & Usman, D. (2013). Research Article Standardization and Its Effects on K-Means Clustering Algorithm. *Research Journal of Applied Sciences, Engineering and Technology*, *6*(17), 3299-3303.
[66]Chiaramonte, L., Croci, E., & Poli, F. (2015). Should we trust the Z-score? Evidence from the European Banking Industry. *Global Finance Journal*, *28*, 111-131.
[67]Skiena, S. S. (2017). *The data science design manual* Cap. 4.3: Z-scores and Normalization. Springer.
Ester, M., Kriegel, H. P., Sander, J., & Xu, X. (1996, August). A density-based algorithm for discovering clusters in large spatial databases with noise. In *kdd* (Vol. 96, No. 34, pp. 226-231).
[68]Kannan, K. S., Manoj, K., & Arumugam, S. (2015). Labeling methods for identifying outliers. *International Journal of Statistics and Systems*, *10*(2), 231-238.
[69]Silver, N. (2012). *The signal and the noise: Why so many predictions fail-but some don't*. Penguin.
[69] Skiena, S. S. (2017). *The data science design manual* Cap. 2.3: Correlation Analysis. Springer.
[70]Hauke, J., & Kossowski, T. (2011). Comparison of values of Pearson's and Spearman's correlation coefficients on the same sets of data. *Quaestiones geographicae*, *30*(2), 87-93.
[71]Wu, Y., Wang, Q., & Shi, Y. (2021). Research on Principal Component Feature Extraction Method Based on Improved Pearson Correlation Coefficient Analysis. In *Advances in Intelligent Information Hiding and Multimedia Signal Processing: Proceeding of the 16th International Conference on IIHMSP in conjunction with the 13th international conference on FITAT, November 5-7, 2020, Ho Chi Minh City, Vietnam, Volume 2* (pp. 82-87). Springer Singapore.
[72] XIE, X. (2019). Principal component analysis. *Wiley interdisciplinary reviews*.
[73]Frutos, E., Galindo, M. P., & Leiva, V. (2014). An interactive biplot implementation in R for modeling genotype-by-environment interaction. *Stochastic Environmental Research and Risk Assessment*, *28*, 1629-1641.
[74]Sarkis, J., & Talluri, S. (2004). Performance based clustering for benchmarking of US airports. *Transportation Research Part A: Policy and Practice*, *38*(5), 329-346.
[75]Wu, W. W. (2011). Beyond Travel & Tourism competitiveness ranking using DEA, GST, ANN and Borda count. *Expert Systems with Applications*, *38*(10), 12974-12982.
[7]6Kaner, C., Bach, J., & Pettichord, B. (2008). Test Run: Group Determination in Software Testing. MSDN Magazine

[77]Ecer, F., Böyükaslan, A., & Hashemkhani Zolfani, S. (2022). Evaluation of cryptocurrencies for investment decisions in the era of Industry 4.0: A borda count-based intuitionistic fuzzy set extensions EDAS-MAIRCA-MARCOS multi-criteria methodology. *Axioms*, *11*(8), 404.
[78]Dai, X., & Kuosmanen, T. (2014). Best-practice benchmarking using clustering methods: Application to energy regulation. *Omega*, *42*(1), 179-188.
West, R. M. (2022). Best practice in statistics: The use of log transformation. *Annals of Clinical Biochemistry*, *59*(3), 162-165.
[79]Camuffo, A., & Grandinetti, R. (2011). I distretti industriali come sistemi locali di innovazione. *Sinergie Italian Journal of Management*, (69), 33-60.
[80]Carloni, M., Ciarrocchi, A., & Micozzi, A. (2020). La vicinanza all'Università? Un'opportunità. Le scelte di localizzazione delle start-up innovative italiane. *L'industria*, *41*(2), 269-289.
[81]Ricciardi, A. (2013). I distretti industriali italiani: recenti tendenze evolutive (Italian industrial districts: recent evolutionary trends). *Sinergie Italian Journal of Management*, *31*(May-Aug), 21-58.

Indice delle figure